Deloitte. | databricks

ENSURING FAIRNESS: *A PILLAR OF TRUSTWORTHY ARTIFICIAL INTELLIGENCE*

Artificial intelligence (AI) has become a key driver of digital transformation across industries. Yet 56% of organizations intend to slow or have already slowed AI adoption, largely because of concerns about emerging risks such as unintended outcomes or biases in model outputs, according to a recent Deloitte study[1].

The success of AI solutions and organizations' willingness to adopt them heavily depends on AI's trustworthiness and consistency.

Many organizations, including the US Department of Defense (DOD), the US Department of Health and Human Services (HHS), and Deloitte have developed like-minded guidelines for using AI. Deloitte's Trustworthy AI™ Framework helps organizations develop ethical safeguards across seven key pillars—a crucial step in managing the risks and capitalizing on the returns associated with artificial intelligence.

DELOITTE'S TRUSTWORTHY AI™ FOR GOVERNMENT & PUBLIC SERVICES

Deloitte's Trustworthy AI™ suite of products and services empower agencies to embrace artificial intelligence while identifying, mitigating and managing AI risk.

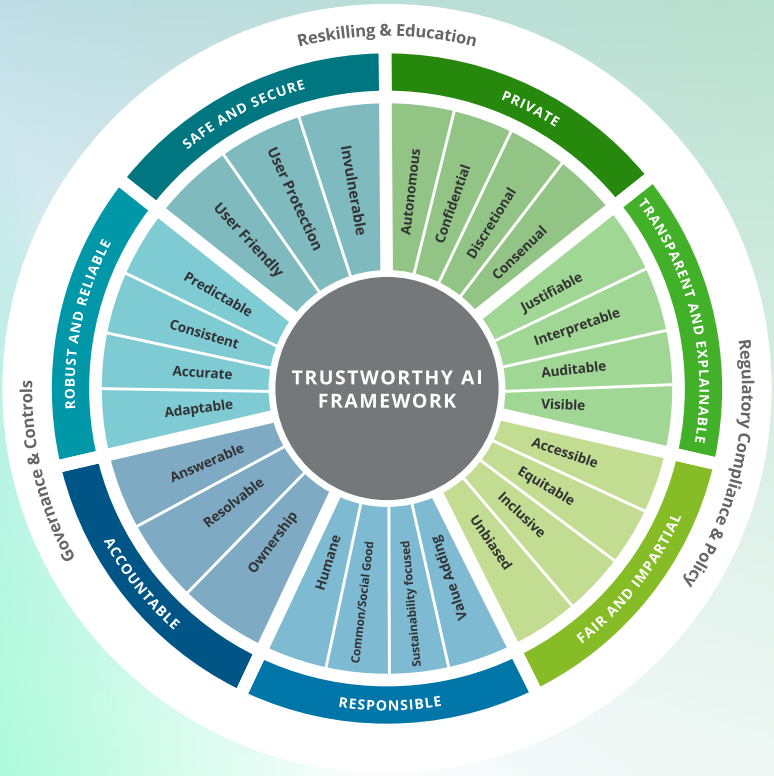


Figure 1:
Deloitte's Trustworthy AI Framework aligns closely with the Department of Defense's Responsible AI framework, the Department of Health and Human Services (HHS) Trustworthy AI Framework, and many other similar initiatives across government and industries.

Safe and Secure

Generative AI (GenAI) systems can be protected from risks (including Cyber) that may cause harm. The models cannot be used as backdoors or to provide harmful, inappropriate, dangerous information.

Robust and Reliable

Consistently produce accurate, coherent, and contextually appropriate responses across a wide range of queries and domains, do not hallucinate, and resilient to erroneous inputs or adversarial accounts.

Accountable

Policies, processes, and controls are in place to determine who is held responsible for all aspects of GenAI systems, including input data and all outputs. There is a system owner who can understand the full breadth of potential privacy and security concerns.

Private

Privacy is respected and personal data is not used beyond its intended and stated use. Prevents unauthorized access to or unintentional disclosure of sensitive information.

Transparent and Explainable

Provide clear insights into GenAI decision-making processes and outputs, enabling users to understand and trust the information generated. Algorithms are open to inspection and can be explained by the responsible owner.

Fair and Impartial

GenAI applications produce fair and unbiased results to all users. They are not influenced by biased data and do not favor certain groups over others.

Responsible

GenAI systems are developed and deployed in an ethical and conscientious manner aligned with social norms, values, and legal regulations.

While each of the seven pillars is crucial to using AI in a way that meets organizations' and individuals' highest standards of trustworthiness, this paper will focus on the Fair and Impartial pillar as an example of Trustworthy AI in action. To help ensure AI remains fair and impartial, the technology must be designed and operated inclusively, with internal and external checks to help ensure equitable application, access, and outcomes.

To understand why AI must remain fair and impartial, let's review an example of how AI may be used in banking to either offer banks and customers an advantage, or lead to discriminatory lending practices.

AI: A CAUTIONARY TALE OR MUTUALLY BENEFICIAL?

By law, financial institutions in the United States must report mortgage lending information to the Consumer Financial Protection Bureau (CFPB). The resulting dataset from the Home Mortgage Disclosure Act (HMDA) includes demographic details about borrowers' race, gender, and age. The data is publicly available and financial institutions can use it to train machine learning models to predict the likelihood of borrowers defaulting on mortgages.

However, without careful attention, this data may lead to biased conclusions because the data alone does not tell the story of unfair housing policies, lending discrimination, redlining, limited access to affordable credit, and other historically inequitable practices in this field.[2] Though lending practices and the demographics of home mortgage applicants today may not look like they did in the past because of historically unjust financial and credit systems, an AI approach that fails to consider these biases could produce discriminatory model outcomes, such as denying loans based on race.

Alternatively, let's consider how AI can help ensure equitable lending practices. Deloitte's Trustworthy AI™ Pipeline (TAP) solution is specifically designed to help ensure the trustworthiness of the machine learning models that underpin AI development. Built on Deloitte's Trustworthy AI Framework, TAP functions primarily through Machine Learning Operations (MLOps) practices. These practices, like DevSecOps (development, security, operations) for software development, focus on developing and deploying machine learning models in a more reliable, scalable, and repeatable manner.

The TAP approach considers how biases can emerge at any point in the machine learning lifecycle, from the data itself to the deployed model, as seen in Figure 2.

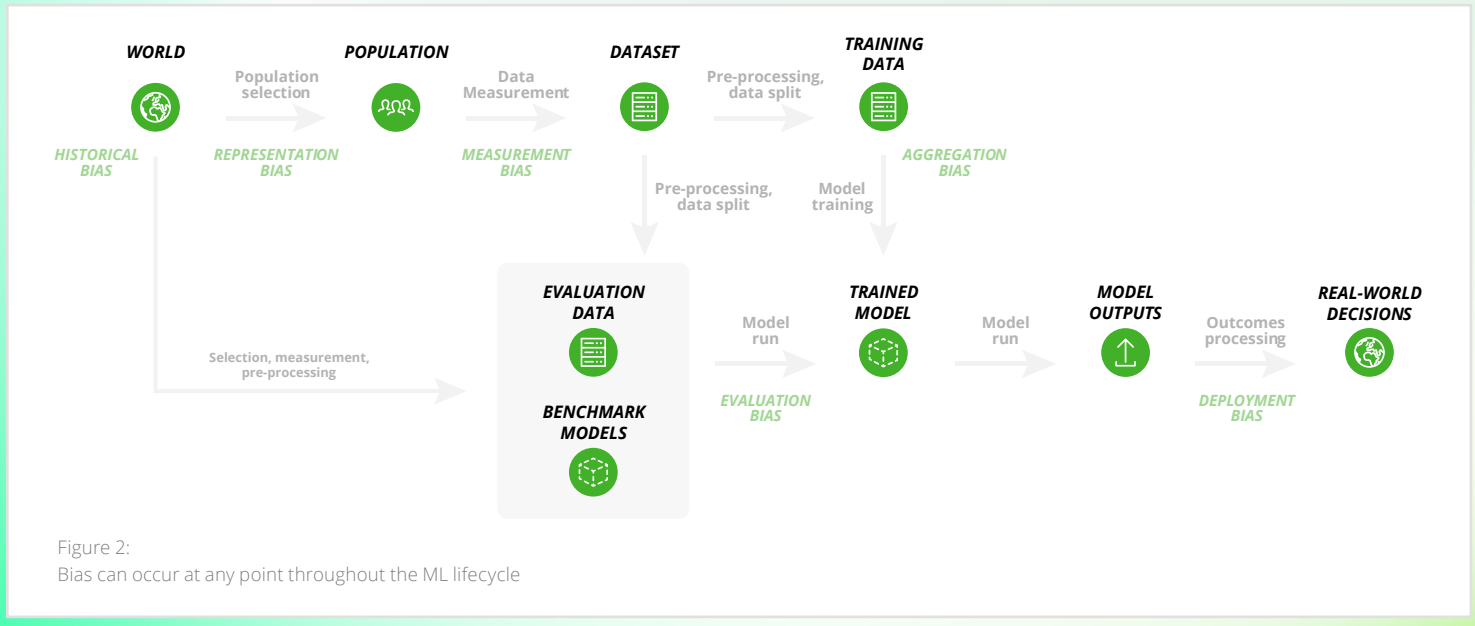
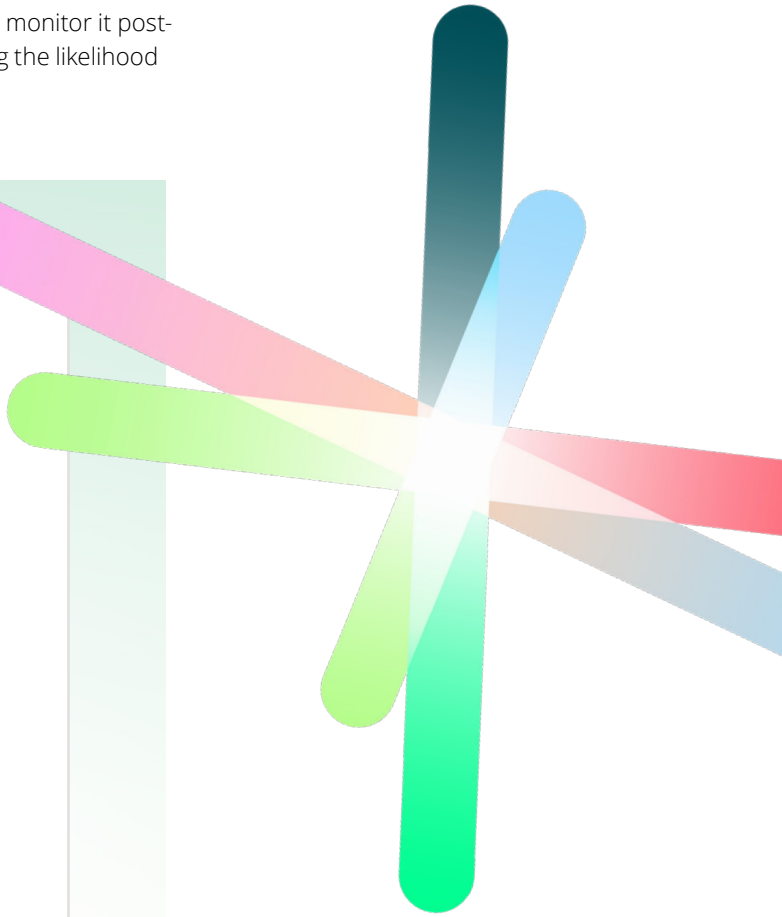


Figure 2:
Bias can occur at any point throughout the ML lifecycle

Using Deloitte's TAP capabilities along with the technology stack from Databricks—a unified platform for data engineering, machine learning, and analytics—organizations can minimize bias throughout the development and execution of an ML mortgage lending model.

In this instance, TAP will analyze the implementation for two main sources of concern—historical bias and representation bias—then measure disparate impact[3] throughout the development process and monitor it post-deployment to maintain an acceptable level of fairness (see Figure 3). With a defined disparate impact ratio (DI) of 0.8 or more, we can measure DI throughout the AI development process and monitor it post-deployment to maintain an acceptable level of fairness when predicting the likelihood of borrowers defaulting on mortgages.



Disparate impact is the ratio of the proportion of positive predictions (y′ = 1) for facet d over the proportion of positive predictions (y′ = 1) for facet a. For example, if the model predictions grant loans to 60% of a middle-aged group (facet a) and 50% other age groups (facet d), then DI = .5/.6 = 0.8, which indicates a positive bias and an adverse impact on the other aged group represented by facet d.

Note: Disparate impact refers to practices in employment, housing, and other areas that adversely affect one group of people of a protected characteristic more than another, and is commonly measured by the 4/5th rule (established by the State of California Fair Employment Practice Commission), which states that if the selection rate for a certain group is less than 80% of that of the group with the highest selection rate, there is adverse impact on that group—though there has been some scrutiny of this benchmark.

Figure 3.
Calculating disparate impact to identify bias and minimize harm

PART 1

ENSURING AI FAIRNESS: EXPERIENCED GUIDANCE

The Deloitte and Databricks alliance combines Deloitte's market-leading industry experience with the Databricks Data Intelligence Platform to solve tough data management challenges and build AI programs to address strategic business objectives. The platform's unified lakehouse architecture combines the attributes of data lakes and warehouses with native governance, AI/ML, and Generative AI features to enable TAP's MLOps tooling with in-depth data preparation, storage, analytics, and modeling.

TAP on Databricks offers a two-step, "metric to measure" governance workflow, which starts before deploying an AI model: Deloitte leverages deep industry experience to work with government agencies to identify and establish the applicable policies or mandates for specific AI use cases. These policies help determine explicit machine learning metrics and thresholds that best fit the use cases and unique mission contexts.

In step two of the "metric to measure" governance workflow, Deloitte leverages their extensive Databricks experience to help organizations configure their Databricks environments and utilize machine learning and engineering to ensure adherence to all seven pillars of Deloitte's Trustworthy AI Framework.

What makes an AI solution uniquely equipped to ensure fairness, or overall trustworthiness?

The answer is deep use case and industry-specific knowledge plus robust technology.

PART 2

ENSURING AI FAIRNESS: FEATURES DESIGNED FOR IMPARTIALITY

The Databricks platform is equipped with dozens of features to empower Trustworthy AI, and this section will focus on those that contribute to the fair and impartial pillar.

Databricks notebooks

In Databricks, notebooks are the primary tool for creating data science and machine learning workflows and collaborating with colleagues. Databricks notebooks provide real-time coauthoring in multiple languages, automatic versioning, and built-in data visualizations.

Additionally, popular trustworthy AI libraries such as SHAP (Shapley Additive Explanations) for explainability and Microsoft's Fairlearn—a toolkit for assessing and improving fairness in AI—can be referenced within Databricks notebooks.

MLflow

Databricks' MLflow tool helps manage the end-to-end lifecycle of machine learning projects. It provides a streamlined workflow for tracking experiments, packaging code, logging parameters and metrics, visualizing results, and managing and deploying models (see Figure 4), all done seamlessly within the Databricks platform.

In the mortgage lending example shared earlier, we explained the importance of a disparate impact ratio and how to calculate it manually; this can be generated in a Databricks notebook and tracked through MLflow.

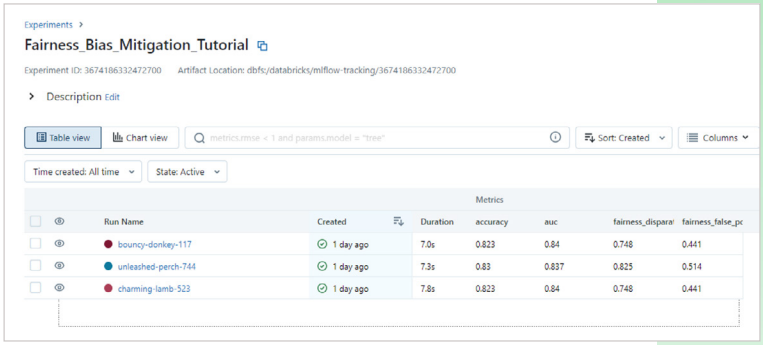


Figure 4:
Databricks' MLflow dashboard for tracking machine learning metrics

MLflow Model Registry

MLflow Model Registry is a central repository where users can store and version their trained machine learning models, allowing them to manage changes to models over time (versions), metadata, and APIs.

MLflow Model Registry also provides dynamic support tracking, which helps organizations easily test for fairness and make adjustments to improve accuracy and impartiality—including increasing or decreasing their DI threshold, tracking other fairness metrics, or updating models to comply with new regulations.

When organizations use Model Registry to deploy models, they ensure only approved and validated models are pushed into production.

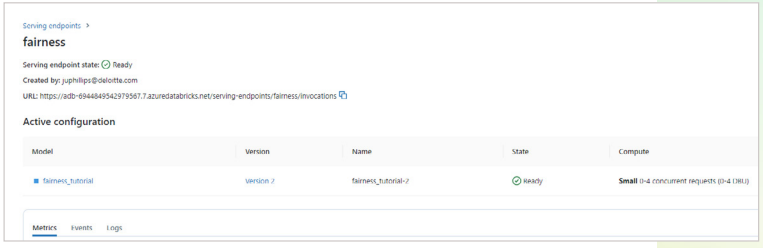


Figure 5:
MLflow serving endpoints view

Databricks pipelines

Databricks pipelines help users define automated workflows that include model versioning and tracking, and improve the governance of the entire machine learning lifecycle from data preparation to deployment.

This is especially critical for a demand-driven market such as home mortgage because—as the composition of the borrower pool changes over time due to macroeconomics or socio-demographic trends—the underlying training data of the model may no longer be representative of the market. This would likely cause a drop in our fairness metric, DI. Automated pipelines allow for rapid retraining of the model using additional, inclusive data to improve fairness despite market data changes.

While tools to automate, track, manage, approve, and deploy AI in a consistent and transparent manner help adhere to fairness thresholds, it's important to note that a model that's considered fair one day, may not be fair the next. Constant monitoring and adjusting is critical to ensuring ongoing fairness.

Unity Catalog

Within the Databricks Data Intelligence Platform, Unity Catalog establishes a central repository for all data assets and models and serves as a precise governance solution for data and AI (see Figure 6). It incorporates AI-powered monitoring and observability to automate error diagnosis, maintain data and ML model quality, and offer proactive alerts for identifying personally identifiable information (PII) data and model drift, thereby preserving data integrity.

In the context of our mortgage lending use case, Unity Catalog helps ensure that lending institutions can identify and rectify biases in their algorithms. For example, if historical lending data indicates biases against applicants of a certain race, Unity Catalog's robust versioning and metadata management features allow data scientists to trace these biases back to specific model versions and associated parameters, so they can correct them.

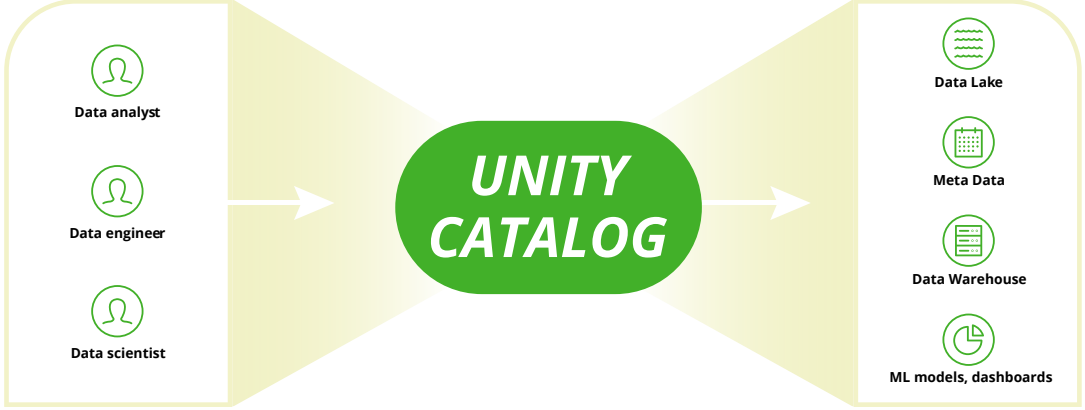


Figure 6.
Unity Catalog acts a central repository for all data assets and models, and serves as a precise governance solution for data and AI.

A FEATURE STORE

is a centralized repository that enables data scientists to find and share features, and ensures that the same code used to compute the feature values is used for model training and inference. The Model Registry and Feature Store seamlessly integrate to simplify secure asset sharing across workspaces and efficient co-administration of both data and AI components. This empowers consistent code application in the computation of feature values, irrespective of the context—be it model training or inference.

DELTA SHARING PROTOCOL

provides a secure channel for data and AI asset sharing, which spans different regions, cloud-based, and platform. This protocol allows easy governance, tracking, and auditing of shared datasets, and facilitates sharing of data assets with suppliers and partners for better coordination of the business while meeting security and compliance needs.

MONITORING

One of Databricks' key offerings is the first-ever unified data and AI monitoring service, which empowers users to simultaneously oversee data quality and AI assets. By continuously monitoring data, organizations can establish quantitative measures to confirm data quality and consistency over time. Any changes in data distribution or model performance are promptly captured and alerted, aiding in identifying potential issues.

DATA QUALITY MONITORING

helps monitor the production model and alerts when a data quality issue happens. Databricks monitoring verifies the incoming data (features, data types, etc.) by checking it against the expected schema and examining the distribution of the input data for any deviation.

MODEL QUALITY MONITORING

helps monitor the quality of the model by comparing predicted values with the actual ground truth labels.

MODEL MONITOR

tracks and evaluates traditional model metrics (F1 score, recall, precision, and accuracy) plus metrics for fairness and explainability. Significant drops in any of these metrics could indicate that the fairness of the model has been impacted and should be reevaluated for alignment with organizational policy.

DATABRICKS LAKEHOUSE MONITORING

tracks various tables within a user's account and monitors performance of machine learning models and endpoints through inference tables. User-friendly features include proactive alerts, quality dashboards to help share insights across the organization, key metric monitoring, and "data drift" prevention to ensure production data doesn't differ from the data used for model testing.

CONCLUSION

As our mortgage lending model use case reveals, organizations across all industries that are deploying AI models should consider the trustworthiness of their solutions before, during, and after deployment. To do so, they need to be able to set reliable Trustworthy AI metrics and measure them over time to reduce risk. And as we've seen with Deloitte's Trustworthy AI Pipeline, "metric to measure" governance workflows and capabilities on the Databricks Data Intelligence Platform allow organizations to set thresholds for fairness, explainability, and privacy, and to ensure model compliance at scale.

Ultimately, this approach can enable a more ideal model governance and level of transparency that facilitates both the inclusion of mission stakeholders and the right documentation for development teams. And when these teams work together, the full value of Deloitte's Trustworthy AI Pipeline, fused with Databricks tooling can be realized. When organizations get easy, go-to workflows within their existing technology stack and more ethical guidelines become the norm, everyone wins.

CONTACT US *TO LEARN MORE TODAY*

Dave Thomas
Databricks GPS Lead Alliance Partner
Deloitte Consulting LLP
davidthomas@deloitte.com

Joe Conti
GPS Trustworthy AI Leader
Deloitte Consulting LLP
joconti@deloitte.com

Bob Stradtman
GPS Risk and Financial Advisory Principal
Deloitte DTBA
rstradtman@deloitte.com

Emily Cole
Databricks Alliance Manager
Deloitte Consulting LLP
emcole@deloitte.com

Shreya Nagesh
GPS Risk and Financial Advisory
Deloitte DTBA
shnagesh@deloitte.com

Richa Vala
GPS AI & Data Engineering
Deloitte Consulting LLP
rivala@deloitte.com

Srijan Karan
GPS Risk and Financial Advisory
Deloitte & Touche LLP
srkaran@deloitte.com

Allie Diehl
GPS Strategy & Analytics
Deloitte Consulting LLP
aldiehl@deloitte.com

Jesse Florig
GPS Strategy & Analytics
Deloitte Consulting LLP
jflorig@deloitte.com

Footnotes

1. https://www2.deloitte.com/content/dam/insights/articles/US144384_CIR-State-of-AI-4th-edition/DI_CIR_State-of-AI-4th-edition.pdf
2. https://www.hud.gov/sites/dfiles/FHEO/documents/AFFH%20Fact%20Sheet.pdf
3. https://www.eeoc.gov/laws/guidance/select-issues-assessing-adverse-impact-software-algorithms-and-artificial#_edn14

About Deloitte

As used in this document, "Deloitte" means Deloitte Consulting, a subsidiary of Deloitte LLP. Please see description of our legal structure. Certain services may not be available to attest clients under the rules and regulations of public accounting.

This publication contains general information only and Deloitte is not, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor. Deloitte shall not be responsible for any loss sustained by any person who relies on this publication.

Copyright © 2025 Deloitte Development.
All rights reserved.