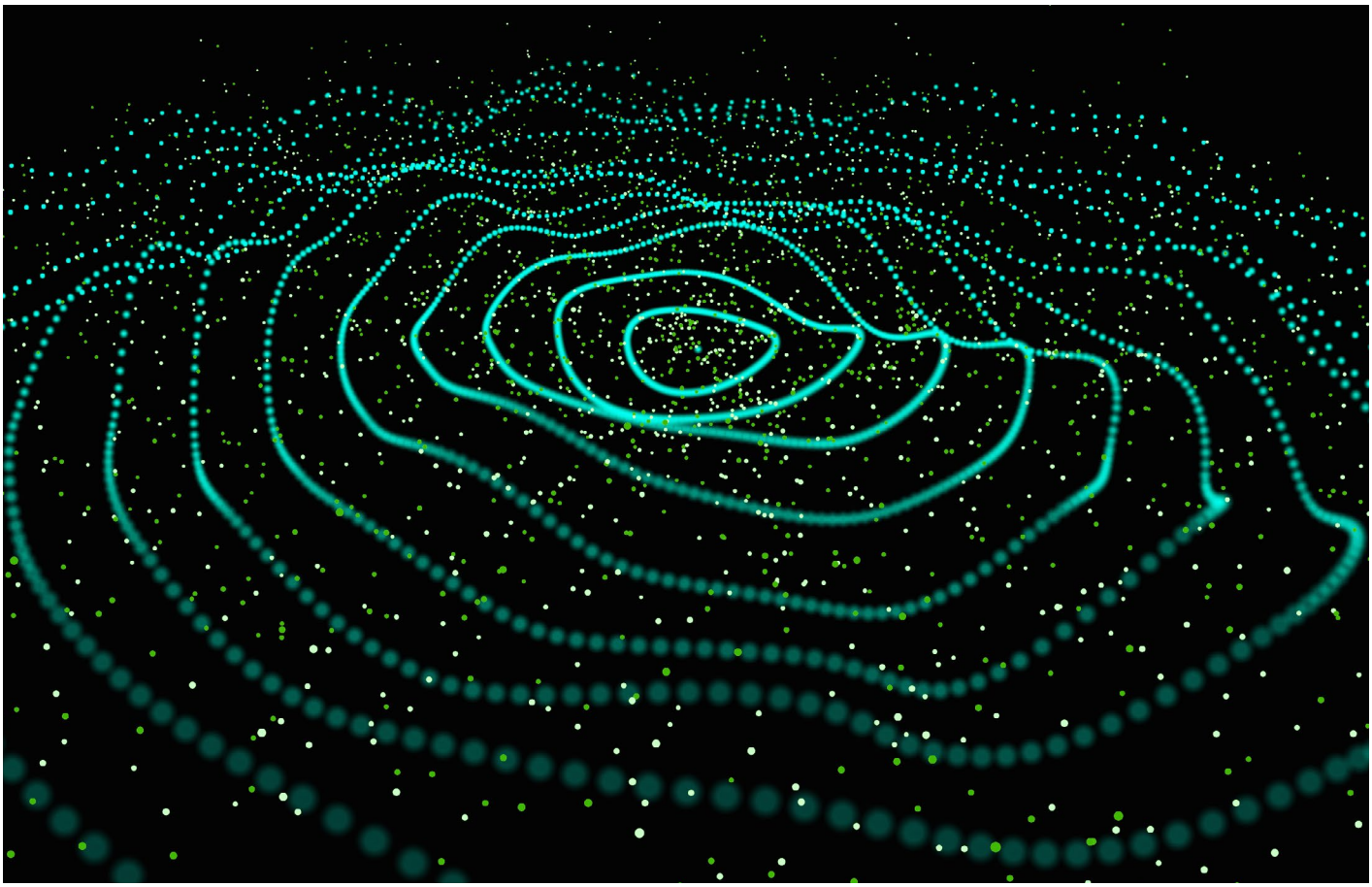

Putting AI Reliability to the Test – Deloitte's AI Qualify Offering for Robust AI

Deloitte's AI evaluation tool AI Qualify assesses the robustness of machine learning models. It subjects the model to a series of tests, examining the resilience, reliability, stability and attack vector vulnerability.

The Need: Building stronger ML models

Every day, Machine Learning (ML) solutions are delivering state-of-the-art results across a wide variety of applications, with accuracy beyond the reach of traditional deterministic models. As ML models increasingly govern critical components of our modern world, their reliability becomes an ever greater concern. This is especially true for ML systems that autonomously take decisions (vs merely advise) – from algorithmic credit scoring to autonomous driving. Despite their general effectiveness, ML-based systems are not immune to failure. It is important to understand their inevitable failure modes, so that we may design ML systems to fail in a predictable and contained fashion, preventing serious harm, protecting against vulnerability to adversaries. We must strive for AI that is as robust and reliable as the traditional systems it enhances or replaces.

Research into adversarial attacks of convolutional neural networks (image recognition) has exposed even sophisticated models to be overly sensitive to a miniscule degree of noise. This opens up opportunities for unscrupulous adversaries who gain access to a model's inner workings to compromise models, to deliberately alter their behavior. Yet all failure modes will not be triggered by sabotage. More commonly, models will slowly lose predictive power as the operational data on which they are applied becomes less representative of the data on which the models were trained – a fundamental concern of MLOps. There is no cure-all approach to fixing these limitations for all ML models; each model must be individually tested and tuned. One thing is clear: building stronger ML models starts with isolating and understanding their weaknesses.

Our Solution: AI Qualify

Ensuring AI is robust & reliable is a central principle of Deloitte's definition of Trustworthy AI. The aiStudio tool "AI Qualify" operationalizes the principle by providing a workbench that tests and verifies the behavior of the ML model under investigation, highlighting existing and potential failure modes.

AI Qualify subjects the model to a series of tests, examining the resilience, reliability, stability and attack vector vulnerability. AI Qualify quantifies model performance along each dimension, capturing it in an individual score as well as an overall measure of robustness. It conducts these analyses at several levels of granularity, providing either a quickly digestible overview or the ability to drill down into targeted areas, as required. AI Qualify also tracks robustness vs predictive power over progressive iterations of model development.

AI Qualify assists model developers identify and address model deficiencies, making for better ML models and a stronger showing of compliance to customer, internal governance and regulatory requirements. It accepts any type of classification or regression model, currently focusing on tabular data, with image and text processing in development.

Advantages/Benefits of Deloitte AI Qualify

- Methodical, structured examination of ML model robustness – a central concern for MLOps
- Confidently apply ML models in critical areas with thorough understanding of failure modes
- Ensure the model complies with regulations
- Save valuable time through automated model evaluation

Example Use Cases

- Designing a model to "fail safely" within a given perimeter
- Assessing degree of model generalization/contextual adaptation
- Testing of edge cases or vulnerability to targeted attack
- Tracking model robustness performance over time (development improvements, model drift deterioration)

Contacts

David Thogmartin
Leader aiStudio
Tel: +49 211 8772 2336
dthogmartin@deloitte.de

Deloitte.

Issue 11/2021