HOW TO BUILD AI YOU CAN TRUST

AND WHY IT MATTERS

PRINCIPAL AUTHORS:

Dr. Joydeep Ghosh
Chief Scientific Officer
CognitiveScale

Matt Sanchez
Chief Technology Officer
CognitiveScale

“By 2022, 30% of organizations will use explainable AI models to build trust.”

Gartner Symposium, 2019

CONTENT

1. WHAT IS TRUSTED AI?

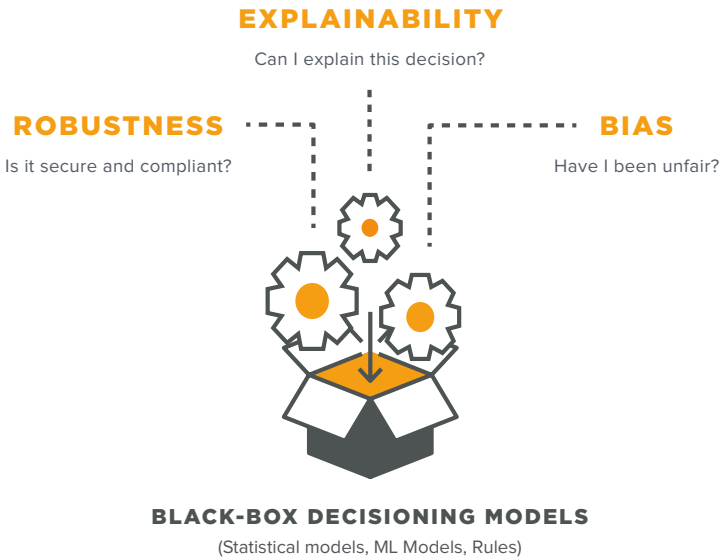
Spending on Artificial Intelligence (AI) is expected to more than double from $35 billion in 2019 to $79 billion in 2022, according to IDC forecasts reflecting the enormous potential societal benefits of AI. Yet broad adoption of AI systems will not come from the benefits alone but from the ability to trust these dynamically evolving digital systems.

Trust is the foundation of all digital systems. Without trust, artificial intelligence and machine learning systems cannot deliver on their potential value. To trust an AI system, humans must have confidence in its decisions. However, AI based automated decisioning systems learn and evolve over time and contain many hidden decision processing layers which can make auditability and traceability challenging.

NEARLY EIGHT OUT OF 10 ENTERPRISE ORGANIZATIONS CURRENTLY ENGAGED IN AI AND ML REPORT THAT PROJECTS HAVE STALLED DUE TO ISSUES OF DATA QUALITY AND MODEL CONFIDENCE

(Dimensional Research, 2019)

This black-box nature of algorithms is creating real business and societal risk about unknown and unintended consequences from rogue AI. This creates the possibility of disastrous repercussions which include the loss of human life if an AI medical algorithm goes wrong, or the compromise of national security if an adversary feeds disinformation to a military AI system. Significant challenges arise from reputational damage, revenue losses, regulatory backlash, criminal investigation, and diminished public trust.



Trusted AI provides framework and software for opening the black-box of AI.

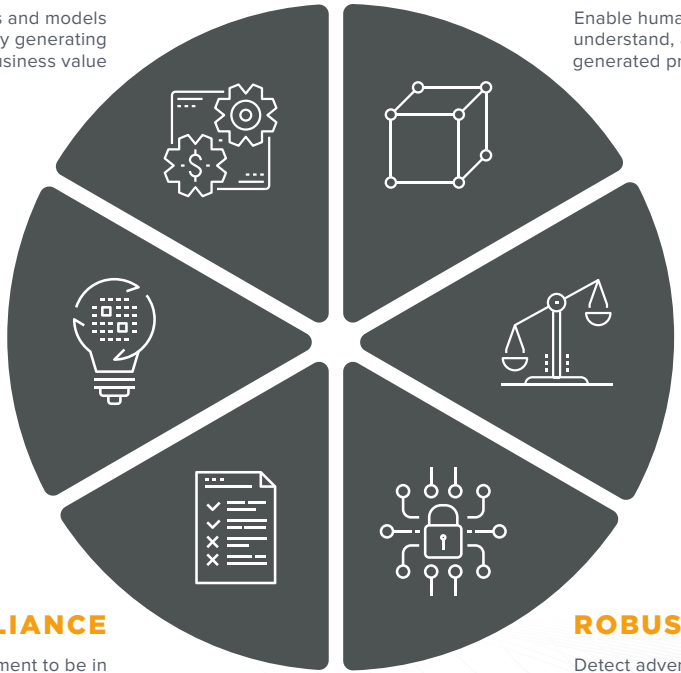
TRUSTED AI HELPS MANAGE AI BUSINESS RISKS

EFFECTIVENESS

Ensure AI systems and models are continually generating optimal business value

EXPLAINABILITY

Enable human users to interpret, understand, and explain machine generated predictions



DATA RISKS

Uncover data drifts, detect data poisoning, and ensure data validity and fit

BIAS AND FAIRNESS

Uncover bias in the underlying data types, data sets, ML model, and AI development process

COMPLIANCE

Align AI development to be in line with current and emerging local and global regulations

ROBUSTNESS

Detect adversarial attacks that disrupt or alter long term model performance

2. HOW DO YOU MEASURE AND MANAGE AI TRUST?

To trust an AI system, we must have confidence in its decisions. We need to know that a decision is reliable and fair, that it can be transparent, and that cannot be tampered with. However, most automated decisioning data and models today — ML algorithms, statistical models, and rules — are black-boxes that often function in oblique, invisible ways for both its developers as well as consumers and regulators.

To help businesses take the first step towards building and maintaining a trustworthy and responsible AI solution, AI Global and CognitiveScale have identified five different types of business risks from automated decisioning systems that need to be managed:

BIAS & FAIRNESS

Trusted AI systems ensure that the data and models being used is representative of the real world and the AI models are free of algorithmic biases to mitigate skewed decision-making and reasoning, resulting in reasoning errors, unintended consequences.

EXPLAINABILITY

AI systems built using Trusted AI principles and software understand stakeholder concerns for decision interpretability and provide business process, algorithmic, and operational transparency so human users will be able to understand and trust decisions.

ROBUSTNESS

As with other technologies, cyber-attacks can penetrate and fool AI systems. Trusted AI systems provide ability to detect and provide protection against adversarial attacks while understanding how issues with data quality impact system performance.

DATA QUALITY

Data is the fuel that powers an AI. AI systems built using Trusted AI principles will ensure user visibility around data drifts, data poisoning, and ensure data validity and fit while ensuring legal justifications to use and process the data.

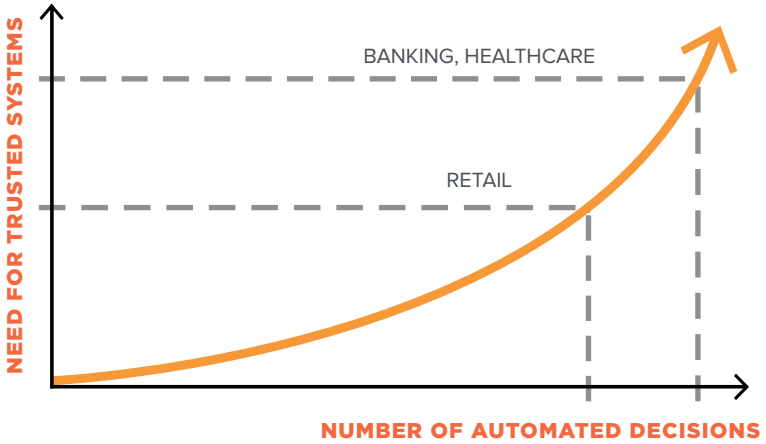
COMPLIANCE

Trusted AI systems take a holistic design, implementation, and governance model that ensures that AI systems operate within the boundaries of local, national and industry regulation and are built and controlled in a compliant and auditable manner.

THE CASE FOR TRUSTED AI

WHAT'S HOLDING AI BACK IN THE ENTERPRISE?

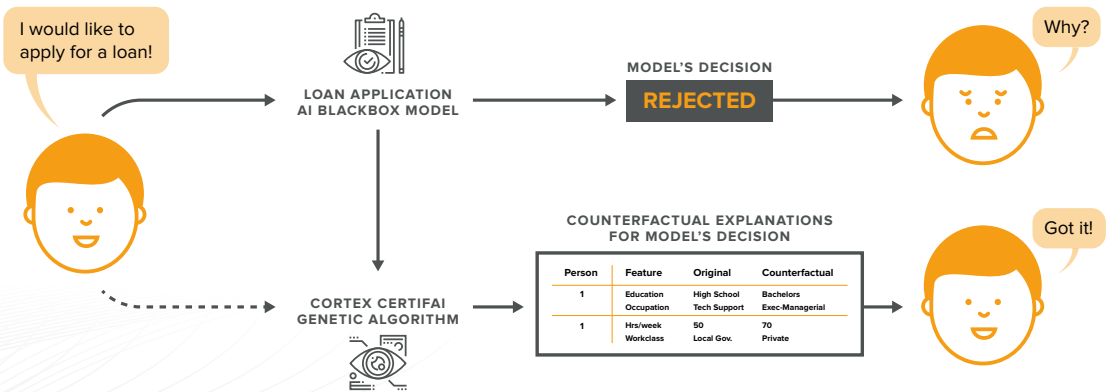
Source: PwC CEO Pulse Survey, 2017



A system built using Trusted AI framework will help answer these questions:

1. How did the system predict what it predicted? (Explanation)
2. If a person got an unfavorable outcome from the model(s), what can they do to change that? (Counterfactuals)
3. Has the AI system or model been unfair to a particular group? (Fairness/Bias)
4. How easily can the model be fooled? (Robustness)
5. Provides results for different business and IT stakeholders (Risk Execs, LOB Owner, Data Scientist/IT)

EXAMPLE: If a user was denied a loan by a machine learning model, an example counterfactual explanation could be: “Had your income been $5,000 greater per year and your credit score been 30 points higher, your loan would be approved.




“By 2022, 40% of employees will consult an AI agent for decision support.”
Gartner Symposium, 2019

4. WHAT IS AN AI TRUST INDEX?

The AI Trust index (ATX) is a numerical score of algorithmic business risk and benefit. It is the first-ever FICO-like composite risk score for any black-box decision making model based on fairness, explainability, robustness, data rights and compliance. The ATX score is a composite of normalized scores along six key dimensions.

1 BIAS & FAIRNESS

The Fairness Score is obtained by comparing the burdens that the model imposes on different segments of the population, using the Gini index, a well-known measure of inequity. Segments can be based on ethnicity, gender, race, age, combination criteria etc. and need to be specified by the user based on domain needs. The "burden" of a group represents the average amount of change required by members of that group to get to favorable outcomes, hence lower numbers indicate more preferential treatment.

2 ROBUSTNESS

Robustness is a measure of how well a model retains a specific outcome given perturbations to the data, whether due to adversarial attacks or natural, statistical variations. Represented as a number between 0 and 100, it is obtained using the Normalized Counterfactual Explanation-based Robustness Score (NCER) Score. A higher number indicates that larger perturbations are needed on average to significantly change a decision, hence indicating more robustness.

3 EXPLAINABILITY

Explainability is also represented as a number 0 and 100, indicating the typical complexity of a counterfactual explanation. The complexity is determined by the number of attributes participating in the explanation for each record in the test data. Counterfactual explanations tend to be shorter for more explainable models. The score for an instance is a non-linear but monotonically decreasing function of the number of attributes that need changing. If only a single change (the minimum possible) is required, the score is 100, and if more than 5 attributes need to change the score is 0.

4 DATA RISK

Represented as a number between 0 and 100, the evaluation of data risks is based on an assessment survey on data management and data privacy practices, as well as statistical tests to evaluate data quality. Work is underway to make this score less survey dependent, and move towards assessment primarily based on the data characteristics, meta data, audit logs, data drift measurements and industry specific (e.g. HIPAA) policies.

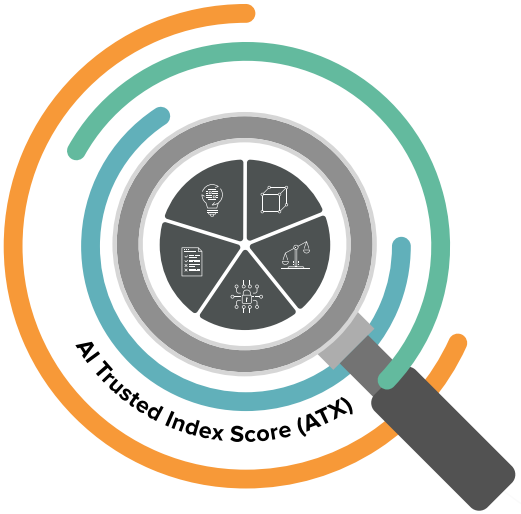
5 COMPLIANCE

Represented as a number between 0 and 100. Primarily based on an assessment survey or directly imported from an organization's Governance, Risk and Compliance (GRC) systems.

6 ACCURACY

Accuracy is a proxy for performance solely in terms of machine learning prediction quality. It denotes an appropriate measure of the accuracy of the predictive model being evaluated for the considered use-case. For classification, this may just be classification accuracy, but could also be defined in terms denote area under the ROC curve (AUROC), or lift at a specified decile. For regression problems, a suitable measure is the (adjusted)-R2 value. For ranking problems, popular metrics include the F1 measure, NDCG, precision, etc.

Trust is the foundation of the digital economy. The AI Trust Index score is based on a statistical analysis of an AI Model's behavior as well as an assessment of the environment in which it is developed and deployed. Specifically, the ATX score captures the risks due to factors such as existence of bias, lack of explainability and transparency, lack of robustness to environmental changes or to data attacks, risks from non-compliance, and risks due to data quality issues.



ATX has been built through an open community collaboration initiative led by AI-Global, a non-profit focused on promoting Responsible and Ethical AI to minimize AI harm.

It provides a simple, common language to help various groups collaboratively build, operate, and evolve high value AI systems. For example:

1. AI and ML engineers use ATX to build and share high quality models and data with built-in AI Trust Index score for explainability, bias and robustness
2. Regulators use ATX to get visibility into how well a company and its management can explain AI decisions to customers, stakeholders
3. Product Managers use ATX to improve confidence in their products' ability to handle model bias and provide end user explainability
4. Brands use ATX to build end user trust through explanations on machine generated decisions
5. CIOs and CDOs use ATX for shared model and data quality representation for interoperability and innovation across their partner ecosystem
6. AI Forensicists use ATX to help provide a means for remediation when AI solutions inflict harm or damages on people or organizations.

For more information visit ai-global.org.


“8 out of 10 enterprise organizations using AI report project stall due to data quality and model confidence.”
Dimensional Research, 2019

5. HOW IS THE AI TRUST INDEX COMPUTED?

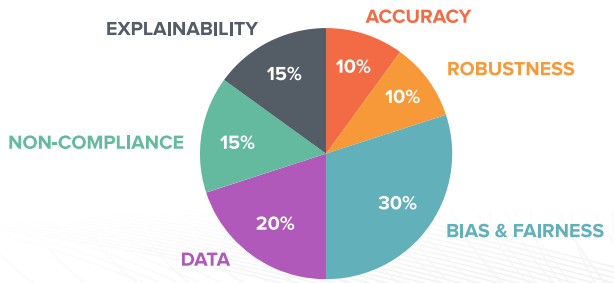
AI TRUST INDEX COMPONENTS

	Grading Approach / Characteristics				
Scoring Component	How Scored	# or Letter	Range Hi	Range Lo	Notes
Robustness	Using Counterfactual determine the score by comparing the distance between two points closer to the boundary	# Absolute Score	100	0	Higher score better
Bias & Fairness	Gini Index	# Absolute Score	100	0	Higher score better
Data Rights	Computed through a combination of metadata inferences and externally through survey questions	# Absolute Score	100	0	Higher score better
Compliance	Externally determined through survey questions or declarations	# Absolute Score	100	0	Higher score better
Explainability	Explanations given out by Certifai app describing hypothetical actions towards getting a better decision from the algorithm	# Absolute Score	100	0	Higher score better
Accuracy	Determined by measurement while training. Provided as a metadata of the AI Model.	# Absolute Score	100	0	Higher score better
Composite (ATX) Score	Computed sum of the above scores after applying relevant weightage to each score.				

AI Trust Index is calculated based on a multi-dimensional analysis of the target black-box models using a unique Counterfactual Explanations based genetic algorithm that does not require access to model internals while providing feasible counterfactual explanations to various roles.

The evaluation criteria, score ranges and their applicability to various industries, business domains, business processes, problem areas and the types of models are expected to evolve and may be guided by Industry / Standardization organizations such as IEEE, AI Global, Academics and leading commercial vendors. The table below illustrates the scoring components and their characteristics.

ATX SCORING COMPONENT WEIGHTS



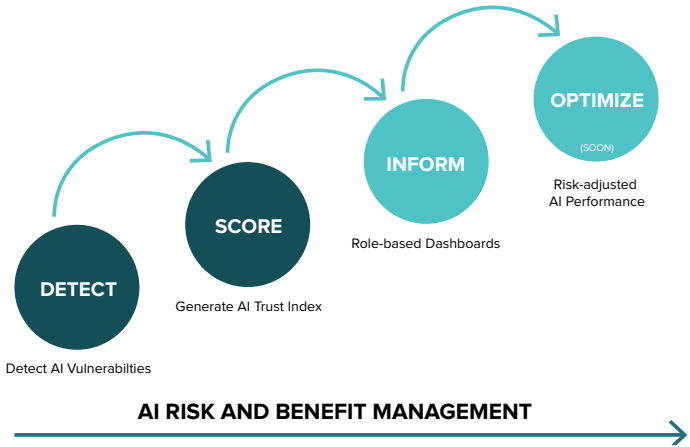
6. WHAT IS CORTEX CERTIFAI™?

AI solutions which learn and evolve over time, and contain many hidden decision processing layers, can make auditability and traceability challenging.

Cortex Certifai is the industry's first AI vulnerability detection and risk management product to address growing customer needs minimizing algorithmic risk while maximizing business and societal benefit. It uses AI to automatically detect and score vulnerabilities in almost all black box models without requiring access to model internals. Cortex Certifai helps enterprises detect and manage in automated decisions systems by answering pressing questions, such as:

1. How did the AI system predict what it predicted?
2. How can a person change an unfavorable outcome?
3. Has the model been unfair to a particular group?
4. How easily can the model be fooled?

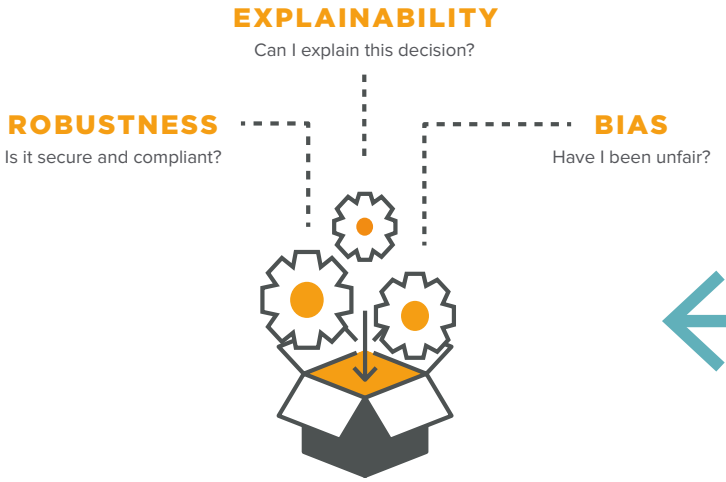
CORTEX CERTIFAI KEY FUNCTIONS



KEY FEATURES

- **Model-agnostic:** Works with any black-box classification model including Rules, Statistical models, ML models
- **Secure:** Does not require access to model internals to detect and score risks
- **Controlled:** Does not involve exchange of models/data across client firewall
- **Extensible:** Allows integration into client's existing Model Governance and GRC work flows
- **Role-based:** Provides personalized results for different IT and business stakeholders

CORTEX CERTIFAI IS THE INDUSTRY'S FIRST AI VULNERABILITY DETECTION AND RISK MANAGEMENT SOFTWARE. IT IS AVAILABLE BOTH AS A STAND-ALONE APPLICATION BEHIND THE FIREWALL OR AS A CONTAINER-BASED KUBERNETES APPLICATION ON ALL MAJOR CLOUDS.



BLACK-BOX DECISIONING MODELS

(Statistical models, ML Models, Rules)

CORTEX CERTIFAI

Automatically detects AI vulnerabilities and generates an unique **AI Trust Index** — without accessing model internals

7. WHAT IS UNIQUE ABOUT CORTEX CERTIFAI™?

CORTEX CERTIFAI PROVIDES THE MOST COMPREHENSIVE COVERAGE OF AI TRUST METRICS

Method	Black-box	Model-Agnostic	Mixed data	Explainability	Fairness	Robustness
CERTIFAI	✓	✓	✓	✓	✓	✓
[Ustun *et al.*, 2019]	✓		✓	✓	✓	
[Watcher *et al.*, 2017]	✓	✓		✓	✓	
[Russell, 2019]	✓		✓	✓	✓	
[Ribeiro *et al.*, 2016]	✓	✓	✓	✓		
[Guidotti *et al.*, 2018a]	✓	✓	✓	✓		
[Carlini and Wagner, 2017]		✓				✓
[Weng *et al.*, 2018]		✓				✓

LAURA
DATA SCIENTIST

PETER
COMPLIANCE OFFICER

MARCUS
BUSINESS OWNER

ADAM
CUSTOMER

SARAH
DATA OWNER

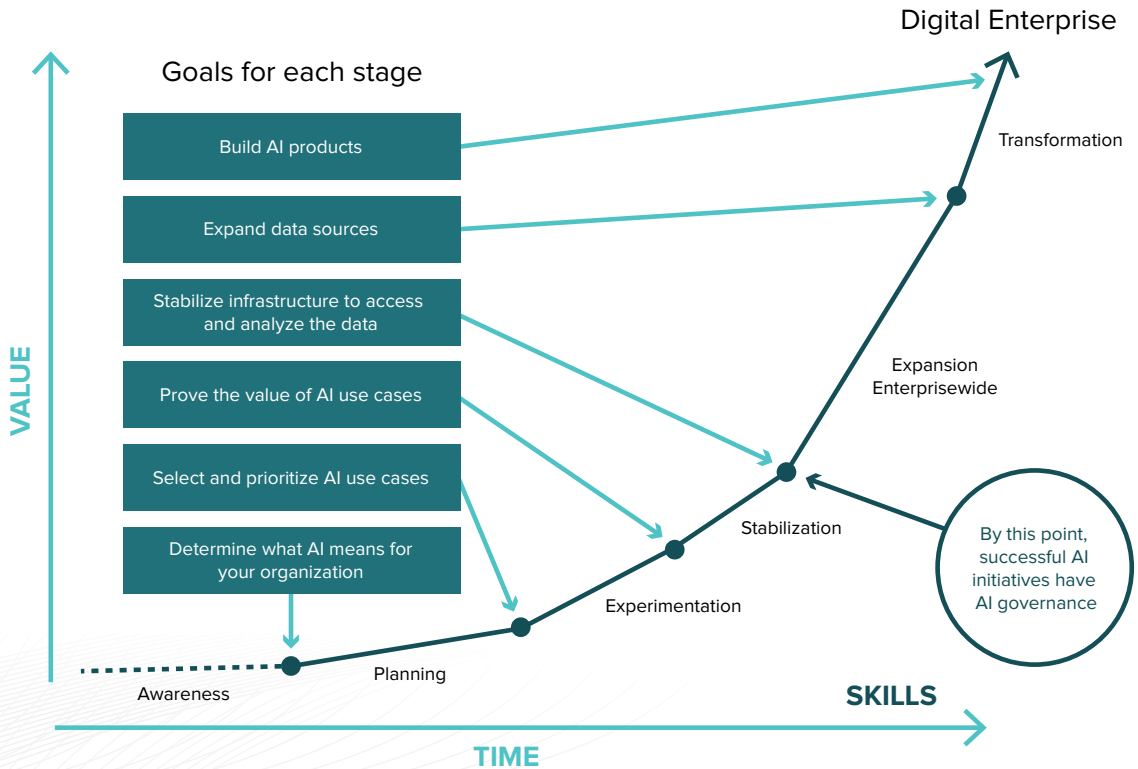
MATT
HEAD OF OPERATIONS

CORTEX CERTIFAI ENGAGES ALL KEY STAKEHOLDERS IN BUILDING TRUSTED AI THROUGH A UNIQUE AI TRUST INDEX PRESENTED THROUGH ROLE-BASED DASHBOARDS.

Cortex Certifai provides the most comprehensive coverage of AI Trust Metrics

8. WHAT ARE THE BEST PRACTICES FOR IMPLEMENTING TRUSTED AI?

TYPICAL STAGES OF AI ADOPTION



9. HOW CAN ORGANIZATIONS GET STARTED?

As AI generates business value and business benefits, it is also giving rise to a host of unwanted, and sometimes serious, consequences. Specifically, lack of explainability and bias management in algorithms is creating significant hurdles in moving AI and ML projects from lab to live. According to a recent survey by Dimensional Research, nearly eight out of 10 enterprise organizations currently engaged in AI and ML report that projects have stalled due to issues of data quality and model confidence.

Much like Security as a Service, Trust as a Service helps accelerate time to value from AI by removing the black-box barrier and driving AI workloads on cloud for generating business value that by PWC's projection is expected to cross $13 Trillion globally by 2030. Not only does an AI Trust Index help businesses adopt and generate value from smarter, quicker automated decisioning systems, it also helps consumers of AI products and predictions build confidence and loyalty in the brands they do business with.

Trust as a Service in the cloud simplifies and accelerates development of automated decisioning and prediction based AI systems that are bias free and transparent in their operations, reflect the company's values, and comply with applicable regulations.

EDUCATE
Demystify AI and its value and risks to businesses

ACTIVATE
Experiment and learn with Certifai Trial version

SCALE
Extend existing GRC processes with Trusted AI

CognitiveScale is an enterprise AI software company with solutions that helps customers win with intelligent, transparent and trusted AI/ML powered digital systems. Our Cortex software and industry AI accelerators enable businesses to rapidly build, operate, and evolve intelligent, transparent, and trusted AI systems on any cloud. The company's award-winning software is being used by global leaders in banking, insurance, healthcare and digital commerce to increase user engagement, improve employee expertise and productivity, and protect brand and digital infrastructure from AI Business risks. Headquartered in Austin, Texas, CognitiveScale has offices in New York, London, and Hyderabad, India, and is funded by Norwest Venture Partners, Intel Capital, IBM Watson, Microsoft Ventures, and USAA.