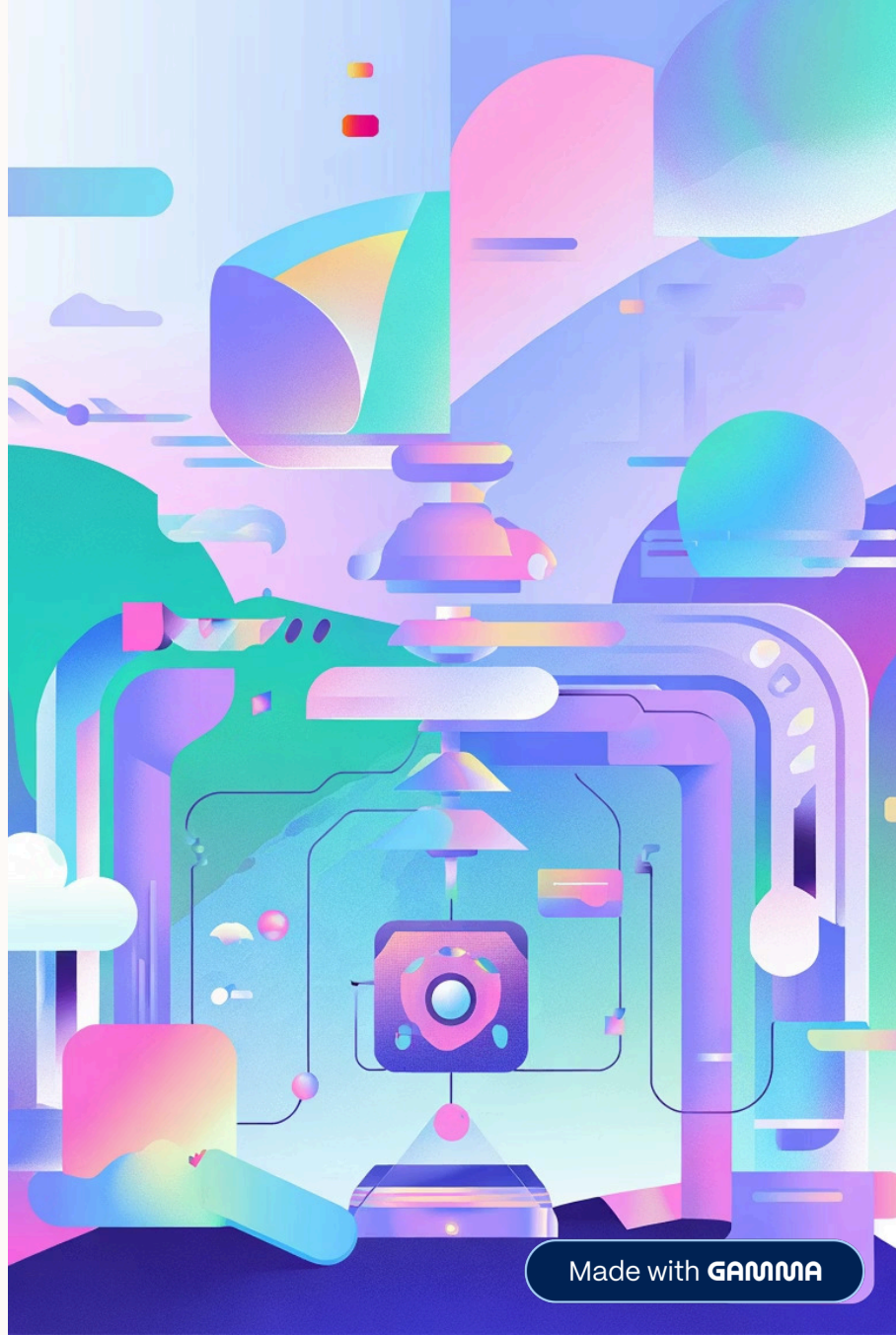
Explainable AI: Designing Domain-Specific Frameworks

Name - Tanishq Kolhatkar **Reg No. - 21MIM10025**

Master's Thesis Review - 1 | AI Governance & Ethics



Thesis Roadmap

This presentation outlines our investigation into Explainable AI (XAI) frameworks that bridge the gap between model accuracy and interpretability. We address the urgent need for transparent AI systems in high-stakes domains where accountability, trust, and regulatory compliance are non-negotiable.

01

Define the Black-Box Problem

Understand opacity, bias, and regulatory risks in deep learning.

02

Validate XAI Solutions

Test bias detection and explainability techniques on real datasets.

03

Build Two Platforms

Enterprise audit tool and clinical decision support system.

04

Demonstrate Impact

Prove feasibility, scalability, and real-world applicability.

The Black-Box Crisis in AI

Deep learning models achieve remarkable accuracy, yet remain fundamentally opaque. This opacity creates three critical vulnerabilities:

Hidden Bias

Discriminatory patterns embedded in training data propagate silently through predictions, affecting hiring, lending, and healthcare decisions.

Accountability Gap

When AI systems fail, organizations cannot explain why—leaving them vulnerable to legal liability and stakeholder distrust.

Regulatory Risk

The EU AI Act, India's DPDP Bill, and emerging legislation mandate explainability. Non-compliance threatens market access and organizational legitimacy.

The Challenge	The Solution	The Outcome
Accuracy without explainability	Explainable AI (XAI)	Trust + Transparency + Compliance

Explainable AI bridges the interpretability-accuracy gap by making model decisions transparent, traceable, and justifiable to stakeholders. Our thesis designs domain-specific XAI frameworks that embed explainability into enterprise governance and clinical practice—transforming AI from a black box into a trusted partner.

Initial Exploration: Proof of Concept

We validated XAI techniques through two foundational experiments using public datasets and established libraries.

Usecases are Explainable Loan Approval , Sentiment Analysis

Experiment 1: Bias Detection

Applied Fairlearn to detect demographic bias in loan approval predictions. Identified systematic disparities across protected attributes and quantified fairness metrics.

- Dataset: UCI Credit dataset
- Tool: Fairlearn bias detection
- Finding: 18% approval disparity

Experiment 2: Model Explainability

Used SHAP and LIME to explain individual predictions in loan approval and sentiment analysis, revealing top contributing features and decision logic.

- Dataset: Public financial + review data
- Tools: SHAP, LIME
- Output: Feature importance rankings

Project 1: X-Trust (Enterprise AI Audit Platform)

X-Trust is an automated compliance engine designed for organizations deploying high-risk AI systems. It provides real-time auditing, bias detection, and transparency reporting—turning governance from reactive to proactive.

Use Case

Enterprise risk and compliance teams audit deployed AI models for bias, robustness, and regulatory adherence.

Core Value

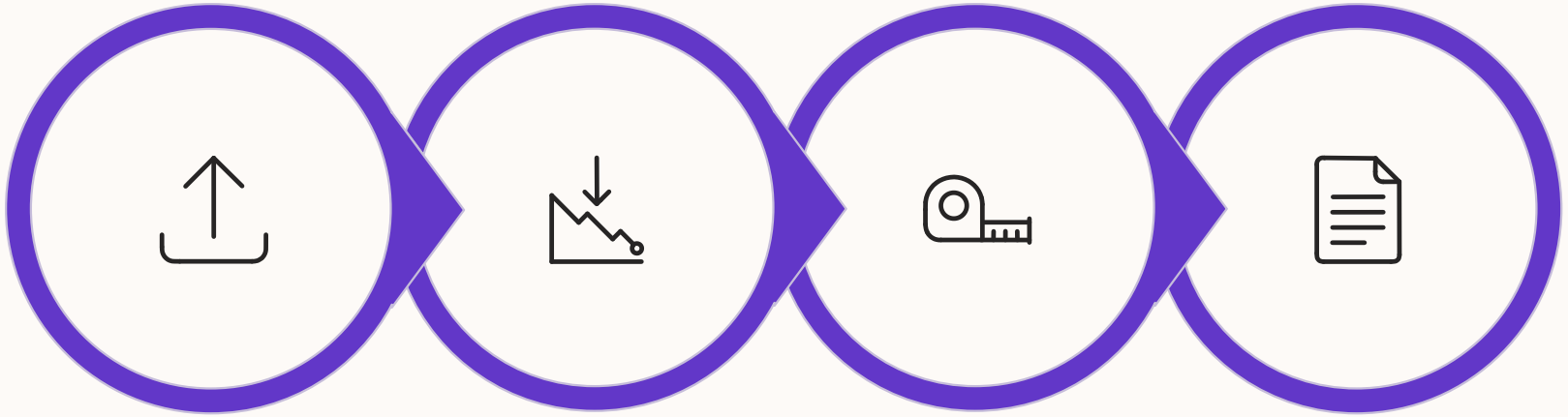
Automated Compliance Check Engine + Model Transparency Report (AI Audit Certificate).

Methods

Bias detection (Fairlearn), explainability (SHAP/LIME), robustness testing, performance monitoring.

X-Trust: Technical Architecture

The platform follows a modular pipeline that transforms raw models into audited, compliant AI systems ready for deployment.



Upload Model | **Automated Audit** | **Evaluation** | **Transpareny Report**

Key Input: Pre-trained ML/DL model. **Key Output:** Comprehensive audit report with fairness scores, explanation dashboard, and compliance certification. **Enterprise Impact:** Reduces audit time from weeks to minutes while providing defensible documentation for regulators.

Project 2: MedExplain (Clinical Decision Support System)

MedExplain brings explainability to clinical AI, enabling physicians to understand and trust AI-generated diagnostic recommendations. It combines prediction accuracy with human-interpretable explanations rooted in patient-specific evidence.

Use Case

Clinicians review AI-predicted disease likelihood alongside transparent, evidence-based explanations for clinical decision-making.

Core Value

Prediction + Top 3 Contributing Factors (SHAP) + Interactive Visual Dashboard.

Societal Impact

Builds physician trust, improves diagnostic accuracy, enables evidence-based care, and ensures patient safety through transparency.

MedExplain: Workflow & Output

The MVP demonstrates how clinical explainability works end-to-end, from patient data ingestion through AI prediction to physician-friendly explanation.

Patient Data Input

Clinical variables (age, blood pressure, cholesterol, glucose levels, medical history).

AI Model Processing

Trained neural network computes disease probability and underlying decision factors.

SHAP Explainability

Identifies top 3 contributing features and their influence on the prediction (positive/negative impact).

Visual Dashboard Output

Prediction score + interactive visualizations (Plotly/Streamlit) showing feature contributions, risk factors, and clinical context.

Result: Physicians receive not just a prediction, but a justified, auditable clinical recommendation backed by patient-specific evidence.

Methodology, Tech Stack & Timeline

XAI Techniques: SHAP (SHapley Additive exPlanations) for global and local interpretability; LIME for model-agnostic local explanations; Fairlearn for bias quantification.

Technology Stack

- **Languages:** Python
- **Libraries:** scikit-learn, TensorFlow/PyTorch
- **XAI Tools:** SHAP, LIME, Fairlearn
- **Visualization:** Plotly, Streamlit
- **Datasets:** UCI Heart Disease, Diabetes, Credit (public)

Thesis 2: Next Steps

- **Q1:** Build X-Trust MVP, integrate audit modules
- **Q2:** Develop MedExplain interface, clinical validation
- **Q3:** End-to-end testing, performance optimization
- **Q4:** Documentation, publication, thesis defense

THANK YOU