

PROBLEM STATEMENT

Track Number: 2

Track Title: Safe, Trusted & Responsible Technology

Video Link: (Optional)

Team ID: _____

Team Name: _____

Institute Name: _____

FairFlow: The RL-Driven Adaptive Bias Firewall

An Enterprise AI Governance Platform for Real-Time Fairness Compliance

TEAM DETAILS

**Member 1 Name****Team Leader**Enrollment Number
Department, Institute Name, KSV**Member 2 Name****Co-Team Leader**Enrollment Number
Department, Institute Name, KSV**Member 3 Name****Team Member**Enrollment Number
Department, Institute Name, KSV**Member 4 Name****Team Member**Enrollment Number
Department, Institute Name, KSV**Guide Name****Guide**

Department, Institute Name, KSV

IDEA DETAILS

Proposed Solution

FairFlow is a "Self-Healing Bias Firewall" that sits between deployed AI models and end-users, ensuring continuous fairness compliance in real-time

Uses **Deep Reinforcement Learning (PPO)** to dynamically adjust decision thresholds, maintaining an optimal balance between Accuracy (Profit) and Fairness (Compliance)

Gatekeeper Agent audits each prediction and decides to **APPROVE**, **DENY**, or **ESCALATE** based on real-time fairness metrics like Demographic Parity and Equalized Odds

Every RL intervention is logged with **SHAP (Shapley Additive Explanations)** for transparent, explainable decision-making

Unique Innovation: Unlike static bias-fixing methods, FairFlow continuously adapts to data drift in production, automatically correcting bias without retraining the base model

TECHNICAL APPROACH

Technologies & Methodology

Technologies Used

Layer	Technology
Base ML Model	XGBoost
RL Agent	Stable-Baselines3 (PPO)
RL Environment	OpenAI Gymnasium
Explainability	SHAP
Backend	FastAPI (Python)
Frontend	Next.js + Recharts
Database	SQLite

Step-by-Step Methodology

1. Data Preparation – Load Adult Census Dataset
2. Bias Simulation – Train biased XGBoost model
3. RL Environment – Custom Gym environment
4. RL Training – PPO with composite reward
5. XAI Integration – SHAP explanations
6. Backend – FastAPI endpoints
7. Dashboard – Real-time React interface

ARCHITECTURE

Proposed Architecture



FEASIBILITY AND VIABILITY

Feasibility & Challenges

Feasibility Analysis

- Technical:** Built using established frameworks (Stable-Baselines3, XGBoost, FastAPI, Next.js)
- Data:** Uses publicly available Adult Census Income Dataset (UCI Repository)
- Resource:** Runs on standard hardware; no GPU required for inference

Challenges & Mitigation

Challenge	Risk	Mitigation Strategy
RL Training Instability	Medium	Pre-trained agent; rule-based fallback
SHAP Latency	Low	Fast mode; cached explanations
Real-time Performance	Medium	Async processing; optimized pipeline
Data Drift Handling	Low	Continuous monitoring; periodic retraining

IMPACT AND BENEFITS

Impact & Benefits

Target Audience Impact

- Banks:** Compliant loan/credit decisions with EU AI Act & GDPR
- HR Firms:** Fair hiring algorithms across demographics
- Insurance:** Equitable premium/claim decisions
- Compliance Officers:** Real-time dashboard & audit trail

Key Benefits

- Regulatory** EU AI Act Article 9 compliance
- Economic** Reduces legal risk & retraining costs
- Social** Equitable AI across all groups
- Operational** Self-healing, continuous monitoring
- Transparency** SHAP explanations for audit trail

COMPARISON WITH EXISTING SYSTEM

Comparison

Feature	Traditional Bias Mitigation	FairFlow (Our Solution)
Approach	Static, one-time fix	Dynamic, real-time adaptation
Data Drift	Requires model retrain	Auto-corrects via RL
Explainability	Limited or none	Full SHAP explanations
Audit Trail	Manual logging	Automatic, immutable log
Integration	Requires model access	Model-agnostic wrapper
Compliance	Periodic manual audits	Continuous monitoring
Human-in-Loop	Not supported	Escalate action available
Deployment	Replace entire model	Plug-and-play middleware