RL Project Phase - I
Bias Mitigation using Reinforcement Learning

Group - 3

Name	**Roll Number**
Aniketh	AM.EN.U4AIE22009
Jatin	AM.EN.U4AIE22024
Kaushik	AM.EN.U4AIE22026

1. Problem Description / Environment Understanding

Loan approval systems often embed historical human biases. The provided dataset (biased_gender_loans.csv) reveals a pronounced disparity: approval rates for men ~42.93%, for women ~18.92%, despite equal distribution in the dataset. This project designs a reinforcement learning (RL) agent capable of learning loan-approval decisions while reducing gender-based disparity.

1.1 Environment Dynamics

The environment simulates a stream of loan applicants sampled from the CSV dataset.
Each timestep presents one applicant with features:
- **salary** (numeric)
- **years_exp** (numeric)
- **sex** (categorical: Man/Woman)

The agent chooses an action — **Approve** or **Reject**.
The environment returns a reward based on:
- correctness of decision using historical label bank_loan, and
- fairness: demographic parity (difference in approval rates between sex groups).

Episodes consist of **100 sequential applicants**. The environment is stochastic because applicants are sampled randomly at each timestep.

1.2 Objectives & Constraints

Main objective:
Maximize loan decision quality (approving "good" applicants and rejecting "bad" ones) while **minimizing demographic disparity** in approval rates.

Constraints:
– No regulatory, budget, or quota constraints provided.
– Only demographic parity fairness is enforced.

2. MDP Components (Model-Free RL Framework)

2.1 State Space (S)

A state represents the applicant presented at the current timestep:

$$s_t = (salary_t,\ years_e xp_t,\ sex_t)$$

State type: continuous + categorical.
State space is **finite but large**, defined by all rows in the CSV dataset.

2.2 Action Space (A)

$$A = (\text{Approve},\ \text{Reject})$$

Binary, discrete, small action set. Suitable for value-based RL (e.g., DQN).

2.3 Transition Probabilities (P)

$$P(s_{t+1}|s_t, a_t)$$

Transitions are independent of the agent's action:
Each next state is sampled from the underlying dataset **IID with replacement**.

Thus:
• Stochastic transitions
• No deterministic linkage between successive states
• Agent influences reward only, not state evolution.

2.4 Reward Function (R)

Reward = **Classification reward** + **Fairness regularization**.

2.4.1 Base classification reward

Since true repayment outcome is unavailable, we must use **historical label** as proxy:

- Approve & historical label = Yes → **+1**
- Approve & historical label = No → **−1**
- Reject → **0**

This avoids fabricating artificial repayment labels and maintains consistency with dataset semantics.

2.4.2 Fairness penalty (demographic parity)

Let:

$$gap = approvalrate(women) - approvalrate(men)$$

Penalty at each timestep:

$$R_{fair} = -\lambda \cdot |gap|$$

Where **λ = 0.5** (moderate fairness pressure).

2.4.3 Final reward

$$R = R_{class} + R_{fair}$$

2.5 Discount Factor (γ)

$$\gamma = 0.99$$

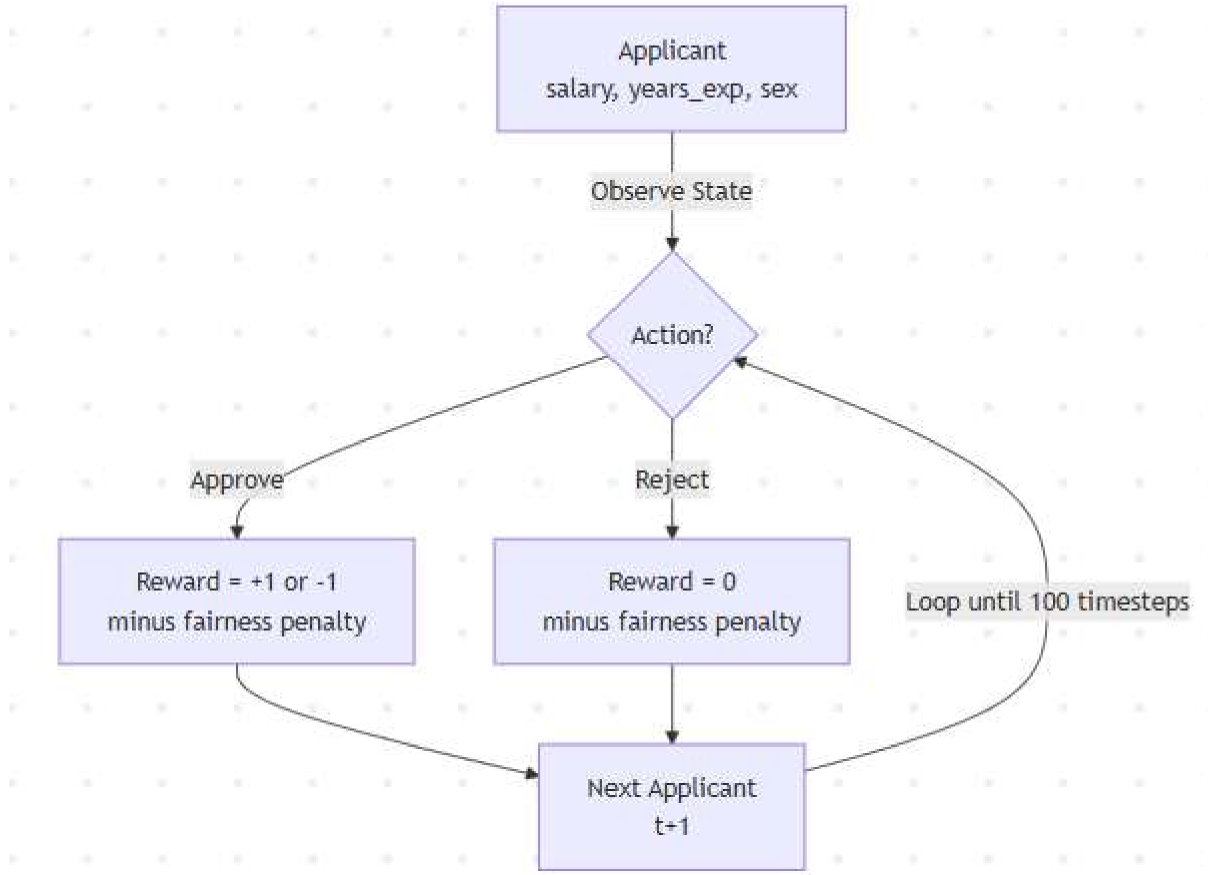
Chosen to encourage long-term fairness and stable policy convergence.

3. MDP Representation

3.1 Sample State–Action–NextState–Reward Table

State (salary, exp, sex)	Action	Historical Label	Next State	Reward
(1700, 25, Woman)	Approve	No	New random applicant	−1 − fairness penalty
(1900, 26, Man)	Reject	Yes	New random applicant	0 − fairness penalty
(1450, 15, Woman)	Approve	Yes	New random applicant	+1 − fairness penalty
(1200, 12, Man)	Reject	No	New random applicant	0 − fairness penalty
(2000, 30, Man)	Approve	Yes	New random applicant	+1 − fairness penalty

3.2 State–Transition Diagram (ASCII Representation)



FP = Fairness Penalty
The environment always transitions to a new applicant until the episode ends.

3.3 Time Step Definition (Δt)

$$\Delta t = \text{One loan applicant decision}$$

4. Objective Formulation

4.1 Optimization Objective

$$\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} \gamma^t R_t\right]$$

Where the optimal policy π balances:
• approval correctness
• fairness (minimizing demographic disparity)

4.2 Episode Termination

Episode ends after **100 applicants**.
Terminal state occurs at **t = 100**.

4.3 Sample Episode Trace

t0: s0 = (1300, 18, Woman), a0 = Reject, R0 = 0 - FP
t1: s1 = (1550, 20, Man), a1 = Approve, R1 = +1 - FP
t2: s2 = (900, 10, Woman), a2 = Approve, R2 = -1 - FP
...
t99: terminal after 100 decisions

The agent experiences fluctuating classification reward and fairness penalties throughout the episode.

5. Methodology (Phase II Plan)

5.1 RL Algorithm Selection

Chosen algorithm: **Deep Q-Network (DQN)**.

Reason:
DQN handles discrete actions, non-linear state spaces, and stochastic sampling efficiently. The dataset has continuous features, making tabular Q-Learning unsuitable. DQN also supports modification of reward shaping for fairness.

5.2 Why DQN is Appropriate (3–4 lines)

DQN handles high-dimensional or continuous state spaces using neural nets instead of lookup tables. The loan-approval environment is stochastic, and fairness penalties introduce non-linear reward structure. DQN offers stability through replay buffers and target networks, making it robust for fairness-aware learning tasks.

5.3 Implementation Plan

Environment:

- Build a Gym-style custom environment that samples applicants randomly.
- Store group-wise approval stats for fairness penalty.
- Implement reward function as defined.

Agent:

- Neural network Q-function approximator
- Experience replay + target network
- ε-greedy exploration

Training:

- 10,000–30,000 environment steps
- Mini-batch training from replay
- Evaluate fairness + accuracy per 100 episodes

Performance Metrics:

- Average cumulative reward
- Approval rate (overall)

- Approval rate by sex
- Statistical Parity Difference

$$SPD = P(approve|Woman) - P(approve|Man)$$

- Disparate Impact (optional)
- Q-value convergence

6. Dataset Observations (For Report Completeness)

10,000 rows; 4 columns.
Sensitive attribute: **sex** (Man/Woman).
Label: **bank_loan** (Yes/No).

Severe Gender Disparity Observed

Approval rate (Men): **42.93%**
Approval rate (Women): **18.92%**

This justifies a fairness-oriented RL approach.

7. Limitations & Assumptions

1. Historical label bank_loan is used as proxy for applicant "quality," although it may itself be biased.
2. True repayment/default outcomes unavailable.
3. Demographic parity fairness is pursued exclusively.
4. Transition model assumes IID sampling independent of action.

8. Conclusion

This Phase-1 submission fully defines the RL environment, formal MDP, reward system incorporating demographic parity, sample dynamics, and the planned DQN-based RL methodology for Phase-II. The environment is grounded in the uploaded biased loan dataset, and all requirements from the assignment PDF have been addressed.