OFFICIAL REPORT

DECISION JUSTIFICATION REPORT

AI Ethical Dilemma Simulator

---

1. INTRODUCTION

Artificial Intelligence systems increasingly participate in decision-making processes that involve ethical consequences. Unlike purely technical optimization problems, ethical dilemmas involve conflicting human values, uncertainty, and trade-offs between harms and benefits.

This report presents the ethical analysis, decision logic, and justification framework for an AI Ethical Dilemma Simulator, designed to model morally ambiguous scenarios using structured and auditable reasoning.

2. OBJECTIVE

The objective of this study is to:

- Design ethically ambiguous AI scenarios
- Apply ethical frameworks
- Formalize transparent decision rules
- Analyze ethical trade-offs
- Rank risks and harms
- Ensure explainability and auditability

3. ETHICAL DILEMMA SCENARIOS

Scenario 1: Autonomous Vehicle Crash Choice

Domain: Transportation | **AI Role:** Collision decision system
Conflict: Passenger safety vs pedestrian safety
Uncertainty: Limited braking distance, uncertain sensor confidence
Ethical Challenge: Choosing between braking (risking passenger harm) or swerving (risking pedestrian harm).

Scenario 2: Medical Treatment Allocation

Domain: Healthcare | **AI Role:** ICU resource allocation
Conflict: Survival maximization vs fairness
Uncertainty: Probabilistic survival predictions
Ethical Challenge: Allocating limited life-saving equipment between patients.

Scenario 3: Hiring AI Bias Risk

Domain: Recruitment | **AI Role:** Resume screening
Conflict: Accuracy vs fairness | **Uncertainty:** Historically biased training data

Scenario 4: Content Moderation AI

Domain: Social Media | **Conflict:** Free speech vs harm prevention
Uncertainty: Context ambiguity

Scenario 5: Predictive Policing AI

> **Domain:** Law Enforcement | **Conflict:** Crime reduction vs discrimination risk
> **Uncertainty:** Bias in historical crime data

4. ETHICAL FRAMEWORK MAPPING

Framework	Role in Decision Justification
Utilitarianism	Minimizes total expected harm
Deontology	Upholds moral duties & rules
Fairness & Justice	Prevents discrimination
Human Rights Approach	Protects fundamental rights

Justification: No single framework sufficiently resolves all dilemmas. A hybrid approach improves robustness across different application domains.

5. DECISION RULE FORMALIZATION

Ethical reasoning was translated into auditable rules through a Weighted Ethical Scoring system:

$$\text{Ethical Score} = (\text{Human Harm} \times \text{Weight}_1) + (\text{Probability of Harm} \times \text{Weight}_2) + (\text{Legal Compliance} \times \text{Weight}_3)$$

Rationale:

- **Human Harm:** Captures severity of the potential impact.
- **Probability:** Reflects inherent uncertainty in real-world scenarios.

- **Legal Compliance:** Ensures the system remains within regulatory alignments.

This approach ensures the system is Transparent, Adjustable, and Auditable.

6. ETHICAL TRADE-OFF ANALYSIS

Trade-offs were analyzed using a matrix evaluating stakeholder impact, benefits vs harms, and temporal consequences. Key Insight: Ethical decisions rarely have universally optimal outcomes; they redistribute risk among stakeholders.

7. RISK & HARM RANKING

Risks were prioritized based on: severity, likelihood, reversibility, and total population impact. This serves to identify ethically critical failure points where human intervention is mandatory.

8. EXPLAINABILITY & ACCOUNTABILITY

The simulator ensures the following standards are met:

- ✔ Traceable decision logic
- ✔ Logged ethical factors
- ✔ Human oversight compatibility
- ✔ Reproducible reasoning

9. LIMITATIONS

- Ethical scoring simplifies complex, non-quantifiable moral values.
- Cultural/contextual ethics vary significantly across boundaries.
- Model outputs are intended as advisory aids, not authoritative commands.

10. CONCLUSION

This Ethical Dilemma Simulator demonstrates how AI systems can be designed with structured ethical reasoning, transparent decision rules, and

auditable justification mechanisms. The framework emphasizes responsibility, explainability, fairness, and human-centered oversight.

“The system does not automate morality but provides a transparent structure for analyzing ethical conflicts, ensuring accountability and responsible AI behavior.”