# AI Systems Reliability and Ethics: A Data Ethics Project

## Background and Context

### The Rise of Public-Facing AI Systems

For this project, students will critically evaluate AI systems by intentionally pushing them beyond normal usage scenarios, documenting exactly how and why systems fail."

The rapid deployment of Large Language Models (LLMs) in public-facing applications has fundamentally changed how organizations interact with customers and process information. What began with simple chatbots has evolved into sophisticated AI systems that:

* Handle customer service inquiries
* Generate content for websites and documentation
* Provide healthcare information and guidance
* Assist with legal and financial advice - Support educational activities
* Process and analyze sensitive data.

Consider these statistics:

* Over 80% of companies plan to implement customer-facing AI by 2025
* AI systems now handle billions of customer interactions daily
* The average person interacts with AI systems multiple times per day, often unknowingly
* Many users place high trust in AI-generated responses, with studies showing 65% of users believe AI systems are generally reliable.

For organizations, the deployment of AI to customers can present ethical and legal issues. The AI is not responsible for incorrect or dangerous answers- the organization is. In this project, you and your group will explore these issues and determine the ethical and legal impact such deployments might have on an organization.

### The Reliability Challenge

Despite rapid deployment, these systems face significant challenges:

1. **Hallucinations** - Generation of false but convincing information - Fabrication of citations and sources - Creation of nonexistent details
2. **Inconsistency**
   * Different answers to identical questions
   * Varying adherence to ethical guidelines
   * Inconsistent performance across contexts
3. **Boundary Issues**
   * Unclear limitations of knowledge
   * Inappropriate confidence in incorrect answers
   * Difficulty with edge cases

### Legal and Ethical Implications

The deployment of potentially unreliable AI systems raises serious concerns:

1. **Legal Liability**
   * Who is responsible for AI-generated misinformation?
   * What duty of care do organizations have?
   * How should damages be assessed?
2. **Ethical Considerations**
   * Impact on vulnerable populations
   * Trust in automated systems
   * Transparency requirements
   * Duty to identify and correct errors
3. **Societal Impact**
   * Spread of misinformation
   * Erosion of trust in institutions
   * Impact on human decision-making
   * Accessibility and fairness issues

## Project Overview

This project examines the reliability and ethical implications of AI systems through systematic testing and analysis. Rather than studying AI capabilities in theory, we will: - Conduct practical experiments to identify failure modes.

The primary goal of this testing is to intentionally induce AI system failures and clearly document the conditions under which they occur to better understand potential risks and ethical implications. This includes the group:

* Documenting reproducibility issues
* Analyzing ethical implications
* Exploring legal and societal consequences

Through this analysis, we’ll explore questions like:

* How reliable are current AI systems?
* What are common failure patterns?
* What are the implications for public deployment?
* How can risks be mitigated?

To do this project, the group needs to pick a fictitious organization as the focus of their research. Here are some potential examples:

* a medical office creating an AI assistant for patients
* a university creating a system to help students register
* a military base using AI systems to help solders know how and when to use their weapons
* a research lab conducting biological studies of Covid-19

How could the issues that you have explored impact these missions? What are the consequences for these failures? Do you think you testing was sufficient for deployment? If not, what level of testing would you need before you recommend the deployment of an AI system for this mission?

## Project Steps

### Step 1: Set Up Project Infrastructure

1. **Create GitHub Repository** (named DATA6550-AI-Reliability)
   * Standard structure:
   * reliability\_report.docx  
     readme.md  
     Experiments/  
      Lastname1/  
      Lastname2/  
      Lastname3/  
      Lastname4/  
     ChatLogs/  
      AI1/  
      AI2/  
      AI3/  
      AI4/  
     Collaboration/  
      WeekA.docx  
      WeekB.docx  
     Analysis/  
      comparison\_matrix.xlsx  
      findings.md
2. **Create shared Word document**
   * Use OneDrive for document sharing
   * Enable track changes under Review tab
   * Share with all team members and instructor

### Step 2: Individual AI System Selection and Planning

The group creates a plan for implementation. Each student selects a unique AI system to analyze. These could be either major commercial systems (ChatGPT, Claude, Gemini) or open source models (Llama, Mistral).

The group should document selection rationale including:

* System capabilities
* Access method
* Known limitations
* Testing approach

### Step 3: Test Design

Your goal is to intentionally probe the system’s weaknesses by crafting queries that challenge its capabilities. Clearly document both the queries and AI responses, noting cases of errors, hallucinations, inconsistencies, or ethical boundary violations.

Design comprehensive test cases across categories:

1. **Factual Accuracy**
   * Historical facts *(e.g., incorrect dates or events)*
   * Scientific knowledge *(e.g., outdated or misunderstood concepts)*
   * Contemporary information *(e.g., incorrect reporting of recent events)*
   * Technical details *(e.g., flawed code examples or calculations)*
2. **Consistency Testing**

* Identical questions across sessions (e.g., test if the AI provides stable answers on repeated trials)
* Variation in phrasing (e.g., slightly rephrase sensitive ethical scenarios to test consistency)
* Context dependency (e.g., changing context mid-conversation to observe reliability)
* Time sensitivity (e.g., asking the same question days apart to test consistency over time)

1. **Boundary Testing**

* Knowledge cutoff dates *(e.g., ask about events occurring after the stated training cutoff date)*
* Specialized knowledge *(e.g., detailed medical or legal queries beyond general training data)*
* Complex reasoning *(e.g., multi-step logic puzzles or ethical dilemmas)*
* Ethical guidelines *(e.g., queries testing adherence to content moderation and sensitive topics)*

1. **Edge Cases**

* Ambiguous queries *(e.g., vague or unclear instructions to test how AI handles uncertainty)*
* Conflicting information *(e.g., providing contradictory context within the conversation to test AI resolution strategies)*
* System limitations *(e.g., queries known to challenge LLM memory or token limits)*
* Policy adherence *(e.g., attempts to get the AI to violate its own stated ethical policies)*

### Step 4: Testing Implementation

Each student conducts systematic testing based on the test design. The logs from the chats will document the results.

During testing, carefully document both your successful attempts (when the AI correctly identifies limitations or responds appropriately) and failures (cases of misinformation, ethical violations, or other errors). Include screenshots or copy text logs to preserve the exact interaction.

### Step 5: Analysis and Documentation

Create a structured analysis covering:

1. **Failure Patterns** -

* Types of errors
* Frequency of occurrence
* Severity assessment
* Reproducibility factors

1. **Ethical Implications**
   * Potential harm scenarios
   * Vulnerable populations
   * Mitigation strategies
   * Deployment considerations
2. **Legal Considerations**
   * Liability issues
   * Regulatory compliance
   * Risk management
   * Documentation requirements

### Step 6: Group Synthesis

Compare findings across systems: - Common failure modes - Unique challenges - Risk patterns - Best practices - Mitigation strategies

### Step 7: Final Deliverables

1. **GitHub Repository:**
   * All chat logs
   * Test documentation
   * Analysis results
   * Meeting summaries
   * Final report
2. **Final Report (5 pages):**
   * Executive Summary
   * Methodology
   * Key Findings
   * Cross-System Analysis
   * Ethical Implications
   * Legal Considerations
   * Recommendations
   * Clearly identify and illustrate significant failure cases found during testing.
   * Provide concrete recommendations for organizational safeguards, ethical training, and transparency practices.
3. **Individual Submissions:**
   * Personal reflection
   * Contribution documentation
   * System-specific findings

## Evaluation Criteria

### Individual Contribution (40%)

* Test design quality
* Documentation thoroughness
* Analysis depth
* Personal reflection
* Clarity and reproducibility of documented test scenarios and interactions.

### Group Deliverables (60%)

* Synthesis quality
* Ethical analysis
* Legal considerations
* Report clarity
* Repository organization
* Clear identification and documentation of significant failure modes and their implications.

## Notes and Guidelines

### Testing Approaches

* Focus on systematic testing
* Document all parameters
* Save complete interaction logs
* Note failed attempts
* Track reproducibility
* Clearly document test conditions (time, prompt wording, context provided).
* Attempt to reproduce failures by repeating problematic queries multiple times and document variations in responses.

### Ethical Considerations

* Consider real-world implications
* Identify vulnerable populations
* Analyze potential harms
* Propose mitigation strategies
* Explicitly test how AI responds to sensitive issues (healthcare, legal advice, vulnerable populations).
* Document any identified biases or inappropriate responses clearly.

### Legal Analysis

* Focus on liability issues
* Consider regulatory requirements
* Document risk factors
* Propose safeguards

### Documentation Tips

* Use consistent formats
* Maintain clear records
* Regular repository updates
* Detailed test logs

## Additional Resources

### AI System Documentation

* System capabilities
* Known limitations
* Usage guidelines
* Terms of service

### Legal References

* AI liability frameworks
* Regulatory requirements
* Industry standards
* Best practices

### Getting Help

* Course forums for technical questions
* Office hours for guidance
* Group discussion for clarification
* Documentation for project structure

## Tips for Success

* Start with simple test cases
* Document everything, including failures
* Consider real-world implications
* Maintain clear communication
* Focus on both technical and ethical aspects
* Consider practical deployment scenarios
* Document assumptions and limitations