CN7050 – Intelligent Systems Coursework

AI BIAS REVIEWER AGENT

Student ID: ____________________
Student Name:
Submission Date: ____________________

1. Application and Workflow Selection

Artificial intelligence systems are now widely employed to generate, summaries, filter, and moderate information across diverse domains including journalism, public policy, recruitment, healthcare, and social media. Although these technologies offer substantial benefits in terms of efficiency, scalability, and automation, they also introduce significant ethical and societal risks—most notably the risk of bias. Bias in AI systems may emerge from multiple sources, such as imbalanced or unrepresentative training data, modelling assumptions, prompt engineering choices, or the reinforcement of historical and structural inequalities embedded within data. When AI-generated content shapes public understanding, influences opinions, or informs decision-making, even subtle forms of bias can contribute to misinformation, exclusion, or the unfair portrayal of individuals and social groups.

Within the news and media domain, the implications of bias are particularly severe. Media content can exhibit framing bias, political bias, cultural bias, or demographic bias through selective emphasis, emotionally charged language, stereotyping, or the omission of critical contextual information. As large language models and automated summarization tools become increasingly integrated into journalistic workflows, there is a growing demand for intelligent systems capable of systematically reviewing textual content and identifying potential bias indicators prior to publication or distribution.

In response to this challenge, this project presents an AI Bias Review Agent developed as a multi-agent intelligent system for analyzing textual content and detecting bias and framing patterns. Rather than attempting to make definitive judgements by classifying content as “biased” or “unbiased,” the system is designed to assist human reviewers by highlighting areas of potential concern and providing structured, explainable analysis. This human-in-the-loop design aligns closely with responsible and trustworthy AI principles, ensuring transparency, accountability, and that final ethical decisions remain firmly under human control.

1.2 Workflow of the Intelligent System

The AI Bias Review Agent is implemented using a modular, agent-based workflow that is coordinated by a central control component. Each agent is assigned a clearly defined and isolated responsibility, which promotes transparency, interpretability, and ease of system maintenance. This agentic decomposition ensures that individual stages of the bias review process can be independently inspected, evaluated, and extended without affecting the overall system integrity. The workflow proceeds sequentially from content ingestion to preprocessing, bias analysis, and structured report generation.

The workflow begins with the input ingestion stage, where the system accepts textual content in multiple formats, including manually entered text, uploaded document files, or predefined sample articles. This flexibility allows the system to be evaluated across a range of realistic usage scenarios. Following ingestion, the preprocessing stage normalizes the text by removing extraneous whitespace, standardizing formatting, and optionally truncating overly long documents to maintain stability and consistency during analysis. These steps reduce noise and ensure that the downstream reasoning process is applied to a clean and manageable representation of the input.

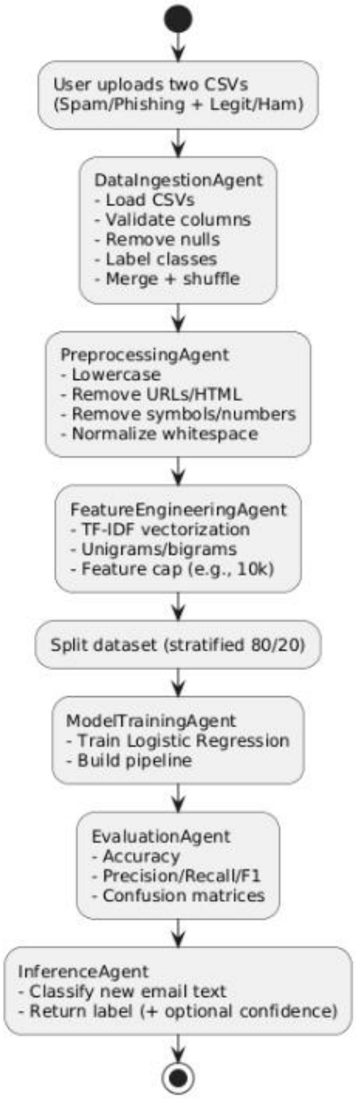


FIGURE 1: WORKFLOW DIAGRAM

1.3 Objectives of the Intelligent System

The AI Bias Review Agent is designed to promote responsible and transparent use of artificial intelligence by systematically analyzing textual content for potential bias and framing risks.

- To support responsible content creation and review by providing a structured and systematic bias analysis framework.
- To identify and highlight potential bias indicators in textual content, including framing effects, imbalanced representation of perspectives, and emotionally charged language.

- To employ an agentic system architecture that decomposes the bias analysis process into specialized agents, ensuring transparency, explainability, and modularity.
- To generate structured, interpretable, and auditable outputs that are suitable for formal review, documentation, and governance processes.
- To demonstrate the applicability of intelligent systems to ethical, regulatory, and AI governance challenges, extending beyond conventional predictive or classification-based tasks.

1.4 Expected Impact

In practical settings, the AI Bias Review Agent has the potential to support journalists, editors, researchers, and compliance teams by significantly reducing the time and effort required for manual bias assessment, while providing consistent and structured analytical support. Rather than replacing human judgement, the system is designed to highlight areas of potential concern and present them in an interpretable manner, thereby assisting reviewers in making informed editorial or governance decisions. This approach reinforces transparency and accountability by ensuring that ethical responsibility remains with human decision-makers.

From an academic perspective, the project demonstrates how agent-based artificial intelligence can be effectively applied to ethical oversight and governance-oriented tasks. By integrating agentic decomposition with qualitative language analysis, the system contributes to ongoing discussions on responsible AI deployment and illustrates how intelligent systems can extend beyond conventional predictive applications to address complex socio-technical challenges.

2. Selection of AI Technologies

Bias detection and analysis have been widely explored across both academic research and industrial practice. Conventional approaches typically rely on fairness toolkits that employ quantitative metrics applied to datasets or model outputs, such as demographic parity, equalized odds, or disparate impact measures. While these techniques are well suited to structured prediction and classification tasks, they offer limited insight into the presence of qualitative bias within natural language content.

In contrast, bias expressed in textual data particularly within news articles and AI-generated narratives—often manifests through framing choices, selective emphasis, tone, and contextual omission. Such forms of bias cannot be adequately captured using numerical fairness metrics alone. Consequently, effective analysis of linguistic bias requires contextual and discourse-level examination, motivating the use of language-aware and interpretative approaches capable of reasoning about meaning, perspective, and narrative structure.

2.1 Identification of Similar Systems

A variety of commercial and research-oriented solutions have been proposed to address bias and fairness concerns in artificial intelligence systems. Prominent platforms such as IBM AI Fairness 360, Google's What-If Tool, and Microsoft Fair learn offer statistical analysis frameworks that evaluate bias by examining dataset characteristics and model outputs. These tools are particularly effective for identifying measurable disparities in structured machine learning tasks, such as classification or risk scoring, where fairness can be assessed through well-defined numerical metrics.

In contrast, bias within the media and journalism domain is commonly assessed through manual editorial review, where journalists and researchers rely on qualitative discourse analysis to identify framing choices, tone, and contextual omissions. While this approach allows for nuanced interpretation, it is inherently time-intensive, subjective, and difficult to scale consistently across large volumes of content. More recently, large language models have emerged as promising tools for qualitative bias detection due to their ability to reason about linguistic patterns, narrative structure, and contextual meaning. When integrated within a human-in-the-loop framework, such models offer the potential to support scalable and interpretable bias analysis while preserving editorial judgement and ethical accountability.

.

2.2 Innovative and Original Approach

The originality of this project is derived from its integration of multi-agent system architecture with large language model–based reasoning for the purpose of bias analysis. Instead of adopting a single, monolithic implementation or relying solely on quantitative fairness metrics, the system decomposes the overall task into a set of specialized and cooperating agents, each responsible for a distinct stage of the workflow, including content ingestion, preprocessing, bias analysis, and structured reporting.

This agentic design closely reflects real-world editorial and governance review processes, where responsibilities are distributed across multiple roles rather than centralized within a single decision-making entity. By enabling each agent's functionality and output to be independently inspected and evaluated, the system enhances transparency, interpretability, and maintainability. Furthermore, this modular approach supports extensibility, allowing additional analytical components or safeguards to be incorporated without disrupting the core system architecture.

2.3 Chosen AI Methods, Technologies and Services

The AI Bias Review Agent is developed in Python and executed within a Jupyter Notebook environment, facilitating reproducible experimentation and interactive system evaluation. This configuration enables users to iteratively test the system, observe intermediate processing stages, and capture execution evidence in a controlled and transparent manner. Such an environment is particularly well suited to academic prototyping, as it supports incremental development and detailed inspection of system behavior.

Text preprocessing is performed using lightweight, rule-based techniques that standardize and normalize input content, thereby reducing linguistic noise and ensuring consistency across analyses. These preprocessing steps help stabilize downstream reasoning and improve the reliability of bias assessment. Bias evaluation itself is conducted using a large language model, which is guided through carefully structured prompts to examine textual content for potential bias indicators and to generate contextual, explanatory reasoning.

To ensure coherent operation, the system incorporates explicit agent coordination logic that governs the flow of information between individual components. This deterministic execution structure enhances transparency, simplifies debugging and maintenance, and supports future extension of the bias analysis pipeline by allowing new agents or analytical stages to be integrated without disrupting existing functionality.

3. Simulation and Development Setup

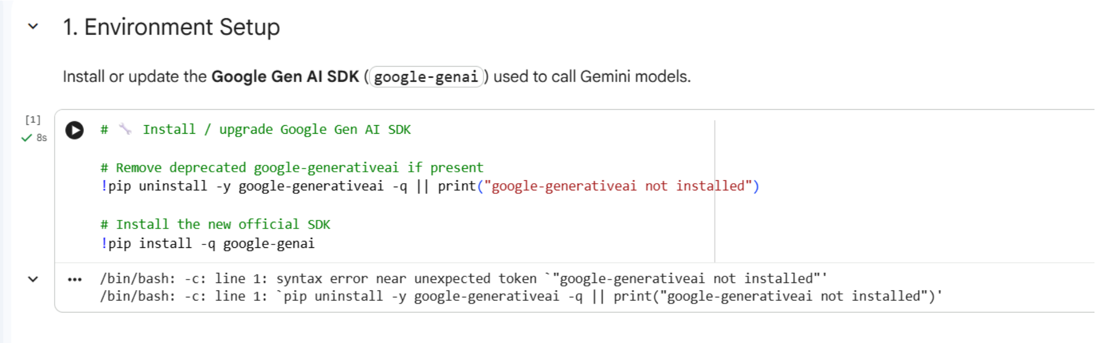
The system is implemented and evaluated as a notebook-based simulation, providing a flexible environment for iterative testing and experimentation. This setup enables users to evaluate the system using a variety of input scenarios, closely observe intermediate processing stages, and capture execution evidence in a structured and transparent manner for assessment purposes. The interactive nature of the notebook supports step-by-step inspection of agent behavior, which is particularly valuable for demonstrating system logic and validating design decisions.

In addition, the notebook-based configuration supports seamless execution on cloud platforms such as Google Colab, ensuring broad accessibility without the need for complex local installation or configuration. This approach enhances reproducibility and portability, allowing the system to be executed consistently across different computing environments while maintaining identical behavior and outputs.

3.1 Simulation Environment and Setup Parameters

Key simulation parameters include the selection of input mode, the maximum length of text analyzed, and the formatting of system outputs. Users are able to provide content through multiple input mechanisms, including manual text entry, file upload, or selection from predefined sample documents, allowing the system to be evaluated across a range of realistic usage scenarios. To maintain stability during analysis and prevent prompt overflow, excessively long documents may be truncated to a configurable length prior to processing.

System outputs are presented in a structured and consistent format, improving readability and supporting systematic evidence collection for evaluation and assessment. This structured presentation also facilitates clearer interpretation of bias indicators and enhances the transparency of the analytical process



1. Environment Setup

Install or update the **Google Gen AI SDK** (`google-genai`) used to call Gemini models.

```
[1]
✓ 8s
# 🔧 Install / upgrade Google Gen AI SDK

# Remove deprecated google-generativeai if present
!pip uninstall -y google-generativeai -q || print("google-generativeai not installed")

# Install the new official SDK
!pip install -q google-genai
```

```
/bin/bash: -c: line 1: syntax error near unexpected token `"google-generativeai not installed"'
/bin/bash: -c: line 1: `pip uninstall -y google-generativeai -q || print("google-generativeai not installed")'
```

FIGURE 2: ENVIORENMENT SETUP

3.2 Agent-wise Task Distribution and Functionality

Agent / Component	Primary Responsibility	Input	Output
Input Agent	Handles ingestion and initial validation of textual content submitted by the user. Ensures that input data is complete and suitable for analysis.	Raw text input (manual entry, file upload, or predefined sample)	Validated text content
Preprocessing Agent	Normalizes and cleans the input text by removing unnecessary whitespace, standardizing formatting, and optionally truncating long documents.	Validated text content	Preprocessed and normalized text
Bias Analysis Agent	Interfaces with a large language model to perform contextual bias evaluation, identifying potential bias indicators such as framing, stereotyping, or imbalanced perspectives.	Preprocessed text	Contextual bias analysis results
Reporting Agent	Aggregates analytical findings and formats them into a structured, human-readable bias review report.	Bias analysis results	Structured bias review report

Coordinating Controller	Orchestrates the execution flow of all agents, ensuring correct sequencing and reliable data transfer between components.	Execution triggers and agent outputs	End-to-end system execution

3.3 Encoding and Data Processing Prior to Simulation

Although the system does not depend on labelled datasets or numerical feature encoding, preprocessing remains a critical component of the analysis pipeline. Text normalization plays an essential role in reducing linguistic noise and ensuring consistency across different inputs, which is particularly important for qualitative bias assessment. By standardizing formatting and structure, the system minimizes variability that could otherwise influence downstream reasoning.

In addition, optional truncation of excessively long inputs is employed to prevent performance degradation and instability during analysis. These preprocessing steps enhance both the reproducibility and reliability of the system by ensuring that bias evaluations are conducted on clean, manageable, and consistently formatted textual representations.

4. Use Case Explanation with Evidence

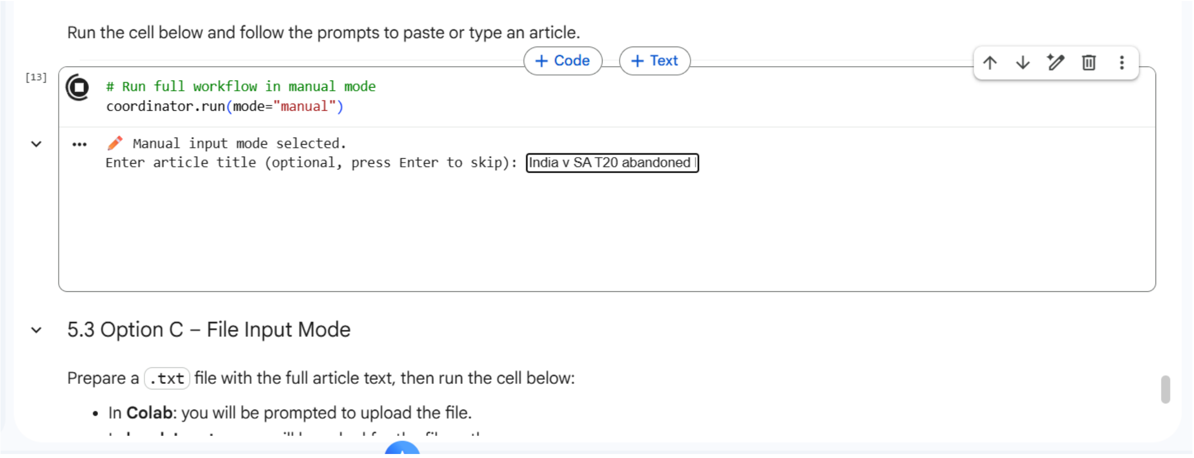
A typical use case involves a user submitting a news article for bias evaluation prior to publication or dissemination. Upon submission, the system ingests the article and applies preprocessing steps to normalize and prepare the text for analysis. The processed content is then examined for potential bias indicators using contextual and linguistic analysis techniques.

The resulting bias review report highlights areas of potential concern, such as framing effects, imbalanced representation of perspectives, or the use of emotionally charged language. By presenting these findings in a structured and interpretable manner, the system supports informed editorial decision-making and enables users to address potential bias issues before content is published.

4.1 Step-by-Step Workflow Execution

First, the user submits the article text through the notebook-based interface. Second, the preprocessing agent cleans and normalizes the input by standardizing formatting and removing unnecessary noise to prepare the text for analysis. Third, the bias analysis agent examines the processed content using a structured prompt, enabling the large language model to identify potential bias indicators and return explanatory findings. Finally, the reporting agent organizes and presents the analytical results in a clear, structured format that is suitable for human inspection and decision-making.

1. **Input Submission:** The user provides the article text through the notebook-based interface.



Run the cell below and follow the prompts to paste or type an article.

```
# Run full workflow in manual mode
coordinator.run(mode="manual")
```

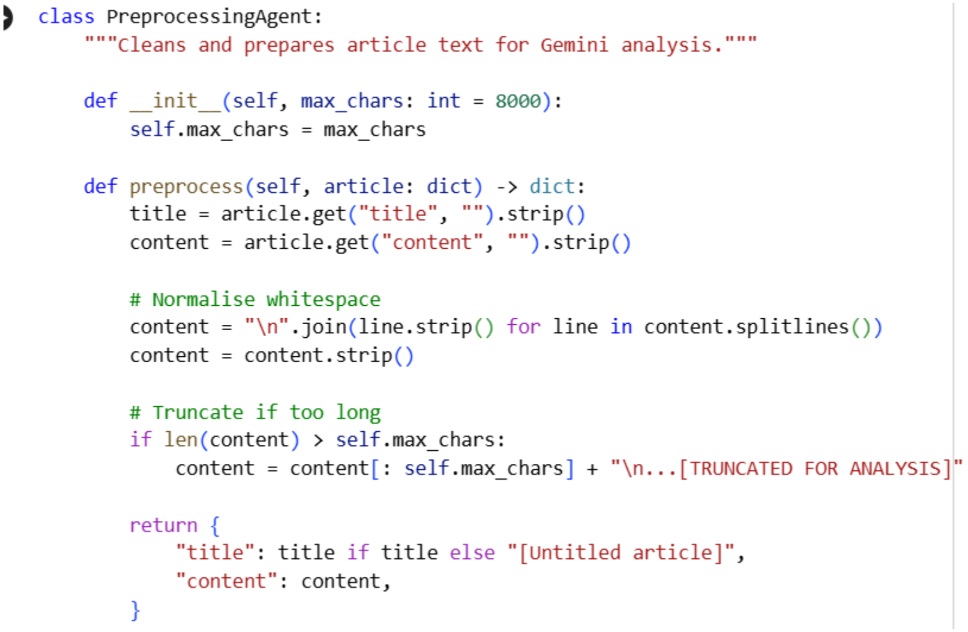
```
✏ Manual input mode selected.
Enter article title (optional, press Enter to skip): India v SA T20 abandoned
```

5.3 Option C – File Input Mode

Prepare a `.txt` file with the full article text, then run the cell below:

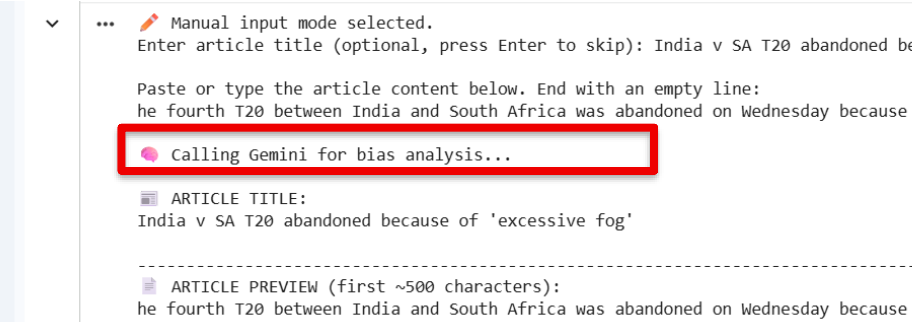
- In **Colab**: you will be prompted to upload the file.

2. **Preprocessing:** The preprocessing agent cleans and normalizes the text by standardizing formatting and removing unnecessary noise to prepare the content for analysis.



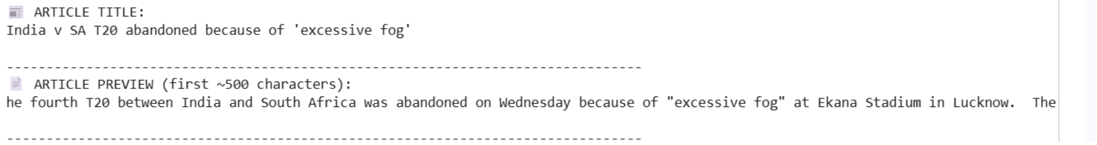
```
class PreprocessingAgent:
    """Cleans and prepares article text for Gemini analysis."""

    def __init__(self, max_chars: int = 8000):
        self.max_chars = max_chars

    def preprocess(self, article: dict) -> dict:
        title = article.get("title", "").strip()
        content = article.get("content", "").strip()

        # Normalise whitespace
        content = "\n".join(line.strip() for line in content.splitlines())
        content = content.strip()

        # Truncate if too long
        if len(content) > self.max_chars:
            content = content[: self.max_chars] + "\n...[TRUNCATED FOR ANALYSIS]"

        return {
            "title": title if title else "[Untitled article]",
            "content": content,
        }
```

3. **Bias Analysis:** The bias analysis agent evaluates the processed text using a structured prompt, enabling the large language model to identify potential bias indicators and generate explanatory findings.



```
✏️ Manual input mode selected.
Enter article title (optional, press Enter to skip): India v SA T20 abandoned be

Paste or type the article content below. End with an empty line:
he fourth T20 between India and South Africa was abandoned on Wednesday because

🧠 Calling Gemini for bias analysis...

📰 ARTICLE TITLE:
India v SA T20 abandoned because of 'excessive fog'

--------------------------------------------------------------------------------
📄 ARTICLE PREVIEW (first ~500 characters):
he fourth T20 between India and South Africa was abandoned on Wednesday because
```

4. **Report Generation:** The reporting agent organizes and presents the analysis results in a clear, structured, and human-readable format suitable for human review and decision-making.



```
📰 ARTICLE TITLE:
India v SA T20 abandoned because of 'excessive fog'

--------------------------------------------------------------------------------
📄 ARTICLE PREVIEW (first ~500 characters):
he fourth T20 between India and South Africa was abandoned on Wednesday because of "excessive fog" at Ekana Stadium in Lucknow.  The

--------------------------------------------------------------------------------
```

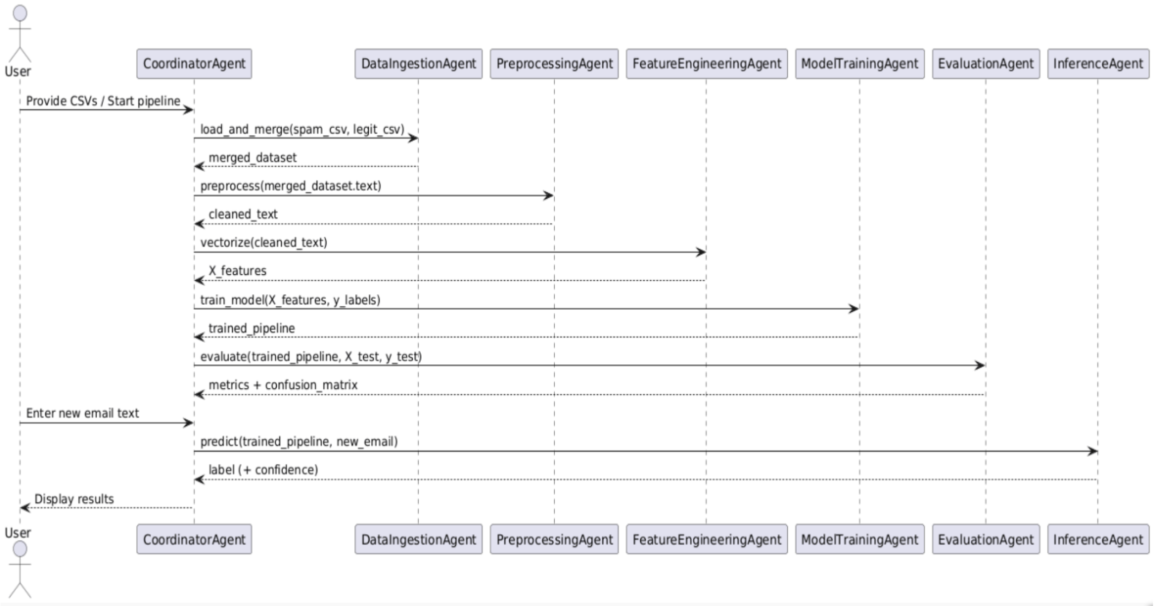


FIGURE 7 : SEQUENCE DIAGRAM

4.2 Developed User Interface and Interaction

The notebook-based interface is intentionally designed to be minimal, prioritizing the demonstration of intelligent system behavior over graphical presentation. Interactive prompts guide users through the processes of content submission, preprocessing, and analysis, enabling clear observation of system logic and agent interactions. This design choice makes the interface well suited for educational demonstration and prototype-level evaluation, while maintaining focus on the underlying intelligent mechanisms rather than front-end complexity.

5. Conclusions

This project demonstrates how multi-agent intelligent systems can be effectively applied to ethical challenges such as bias detection in textual content. By decomposing the overall workflow into specialized and cooperating agents, the system achieves a high level of transparency, modularity, and interpretability. The integration of large language model–based reasoning enables the AI Bias Review Agent to analyses complex linguistic patterns, framing choices, and contextual cues that are difficult to capture using purely quantitative methods. Importantly, the system is designed to support, rather than replace, human judgement, ensuring that ethical responsibility and final decision-making remain firmly under human control.

From a broader perspective, the project highlights the potential of agentic AI architectures for responsible and trustworthy AI deployment. By making each analytical stage explicit and inspectable, the system aligns with key principles of explainability, accountability, and governance. This approach demonstrates how intelligent systems can move beyond traditional predictive or classification-based applications to address nuanced socio-technical problems that require contextual understanding and ethical sensitivity.

Several directions for future work are identified. Quantitative bias metrics could be integrated alongside qualitative analysis to provide a more comprehensive assessment framework. The system could also be extended to support multimodal bias detection, incorporating analysis of images, videos, or audio content commonly used in digital media. In addition, deploying the system as a scalable API service would enable integration into real-world editorial, regulatory, or compliance workflows. Finally, further research could focus on benchmarking system outputs against human-annotated bias datasets to assess consistency, reliability, and alignment with expert judgement, thereby strengthening the system's validity and practical applicability.

REFRENCES

- **Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021).**
  A survey on bias and fairness in machine learning.
  ACM Computing Surveys, 54(6), 1–35.

- **Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021).**
  On the dangers of stochastic parrots: Can language models be too big?

Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT).

- **Wooldridge, M. (2009).**
  An introduction to multi-agent systems (2nd ed.).
  John Wiley & Sons.

- **Manning, C. D., Raghavan, P., & Schütze, H. (2008).**
  Introduction to information retrieval.
  Cambridge University Press.

- **Floridi, L., Cowls, J., Beltrametti, M., et al. (2018).**
  AI4People—An ethical framework for a good AI society.
  Minds and Machines, 28(4), 689–707.