# 24RM801 RESEARCH METHODOLOGY FOR ENGINEERING

Assignment 2:
Identifying and Mitigating Research Bias

Samrat Kar
bl.sc.r4cse24007

12th Jan 2025

# Bias identification

1. **Dataset Bias**

a. Definition: Bias introduced by imbalanced or incomplete datasets.

b. Examples:

i. Overrepresentation of Western aviation scenarios in training data, leading to underperformance in scenarios involving other regulatory environments (e.g., ICAO vs. FAA differences).

ii. Lack of diverse weather, terrain, or traffic scenarios in training data.

c. Impact:

i. Errors in rare but critical situations, such as emergency landings in specific geographic regions.

**2. Contextual Misinterpretation Bias**

a. Definition: LLMs may fail to understand the context in highly specialized domains like aviation.

b. Examples:

i. Misinterpreting ambiguous phrases in pilot-controller communications, such as "ready for departure" vs. "requesting takeoff clearance."

ii. Generating irrelevant suggestions for non-standard phrases during emergencies.

c. Impact:

i. Miscommunication between the model and human operators during critical moments.

**3. Operational Environment Bias**

a. Definition: Models may not generalize well to diverse operational environments.

b. Examples:

i. Underperformance in non-English-speaking regions or in scenarios involving multi-lingual crews.

ii. Misunderstanding cross-cultural variations in decision-making.

c. Impact:

i. Reduced trust from non-native English speakers or crews trained in different regulatory systems.

**4. Interface and Usability Bias**

a. Definition: The way the model's outputs are presented may unintentionally bias pilot decisions.

b. Examples:

i. Recommendations that are overconfidently presented, leading to reduced pilot autonomy.

ii. Insufficiently visualized uncertainty metrics.

c. Impact:

i. Over-trust or mistrust in the system.

**5. Temporal Bias**

a. Definition: Models trained on static data may fail to adapt to evolving operational protocols or standards.

b. Examples:

i. Incorrectly handling changes in ICAO phraseology standards or evolving safety protocols (e.g., changes in runway naming conventions or SIDs/STARs procedures).

c. Impact:

i. Providing outdated or misleading guidance during operations.

**6. Confirmation Bias in Interpretability Tools**

- Definition: Visualization tools (e.g., attention maps) may highlight patterns that confirm user expectations rather than uncovering true causality.
- Examples:

i. Attention weights disproportionately focusing on inputs with high linguistic salience but not operational relevance (e.g., "landing" over "ILS frequency").

- Impact:

ii. Over-reliance on spurious correlations instead of actionable insights.

# Bias Mitigation

1. **Dataset Diversification**
    a. Approach:
        i. Curate datasets that cover diverse geographic regions, weather conditions, regulatory environments, and emergency scenarios.
        ii. Include historical aviation incidents and their transcripts to improve robustness in critical situations.
    b. Implementation:
        i. Use synthetic data generation to augment underrepresented scenarios (e.g., rare weather conditions or high-altitude airports).
    c. Regularly update datasets to incorporate changes in aviation standards.
2. **Context-Aware Training**
    a. Approach:
        i. Incorporate domain-specific knowledge into the model.
        ii. Use fine-tuning with aviation-specific corpora, including ATC transcripts, NOTAMs, and operational manuals.
    b. Implementation:
        i. Train models with structured dialogues annotated for context (e.g., intention, command, response).
        ii. Use hierarchical modeling techniques to better understand the structure of pilot-controller communications
3. **Fairness and Robustness Testing**
    a. Approach:
        i. Evaluate the model under diverse scenarios to ensure fairness and robustness.
    b. Implementation:
        i. Use adversarial testing to uncover biases and spurious correlations.
        ii. Develop benchmarks specific to flight deck applications, such as emergency scenarios, multi-lingual interactions, and non-standard operations.
4. **Visualization and Feedback Mechanisms**
    a. Approach:
        i. Design interpretable dashboards that present model outputs with clear uncertainty metrics.
    b. Implementation:
        i. Visualize key decision pathways (e.g., attention across layers) in real-time during operations.
        ii. Provide confidence scores for each recommendation and flag ambiguous outputs for human review.

3. **Temporal Adaptation**

Approach:

Periodically retrain or fine-tune models with updated datasets to adapt to evolving standards.

Implement mechanisms for online learning in deployment.

Implementation:

Use active learning frameworks where users provide feedback on model outputs to refine future predictions.

4. **Causal Interpretability Tools**

Approach:

Move beyond attention maps to causal attribution methods.

Implementation:

Apply Structural Causal Models (SCMs) to map the relationship between input features (e.g., commands, environmental data) and outputs (e.g., recommendations).

Use counterfactual analysis to test the robustness of outputs in varying scenarios.

5. **Multilingual and Cultural Bias Mitigation**

Approach:

Train models on multi-lingual datasets and cultural-specific aviation corpora.

Implementation:

Incorporate translation tools fine-tuned for aviation-specific language.

Use reinforcement learning with human feedback (RLHF) from diverse global pilots and ATC professionals.

# Biases in reseach and their Mitigation

Researcher bias plays a significant role in shaping research findings, as it can influence the design, methodology, data interpretation, and even the conclusions drawn from the study. Bias can manifest in various forms, including confirmation bias, where researchers unintentionally favor data or interpretations that support their hypotheses, or selection bias, where certain data is chosen while ignoring others. Such biases are not only a result of personal beliefs or preferences but can also stem from unconscious factors, institutional pressures, or the researcher’s cultural background. These biases can skew results and lead to conclusions that do not accurately reflect the broader context or reality.

To mitigate the impact of researcher bias, transparency is crucial. Being transparent about the research process—such as the choice of methodology, data sources, assumptions, and limitations—allows others in the academic community to critically evaluate the findings. Open access to datasets, code, and analytical methods fosters a culture of reproducibility, ensuring that the results are not only verifiable but also scrutinizable. Transparency also helps in identifying and addressing potential sources of bias, ensuring a more objective and balanced approach to research. Furthermore, disclosing any conflicts of interest and acknowledging the limitations of a study increases credibility and allows for a more accurate assessment of the research's implications. In the context of aviation, for instance, where human safety is involved, transparency and addressing biases become even more critical, as biased findings can have direct consequences on operational practices and safety protocols. Therefore, researchers must prioritize honesty, openness, and objectivity to foster a reliable and ethical research environment.