FT2025
SWEDISH AEROSPACE AT THE CROSSROADS
The 12th Swedish Aerospace Technology Congress
FT2025 in Stockholm
October 14-15, 2025

Improving Decision-Making of Civilian and Military Pilots Under Pressure An Artificial Intelligence-Based Approach Using Prospect Theory

Giovane De Morais[1], Carolina Leão Giollo[2], Moacyr Machado Cardoso Júnior[1], and Emilia Villani[1]

[1]*Manufacturing Competence Center, Instituto Tecnológico de Aeronáutica (ITA), Brazil*

E-mail:, moacyr@ita.br, giovane@ita.br, evillani@ita.br

[2]*Department of Technology, Centro Universitário Adventista de São Paulo (UNASP), Brazil*

E-mail:, carolgiollo@gmail.com

Abstract

Large Language Models (LLMs) are entering safety-critical domains, yet their alignment with expert judgement under risk remains uncertain. We compare decision patterns among professional pilots (N=36), lay participants (N=36), and a local LLM (single-case, N=1) across six hypothetical scenarios spanning operational hazards, an ethical dilemma, and two canonical Prospect Theory framings. In loss/gain frames, humans and the LLM showed biases consistent with Prospect Theory. However, in operational scenarios, pilots consistently chose the safest option—reflecting recognition-primed pattern matching and safety culture—whereas the LLM selected risk-seeking, utilitarian strategies. In the ethical dilemma, nearly all humans favored a duty-of-care decision while the LLM maximized aggregate outcomes. These results indicate that surface-level mimicry in abstract tasks does not translate to context-grounded, duty-constrained reasoning required in aviation. We outline a multilayer alignment agenda (expert RLHF, constitutional duty constraints, formalized taboos, domain red-teaming) to reduce misalignment before deployment in safety-critical settings.

Keywords: Large Language Models, Aviation Safety, Prospect Theory, Human-AI Alignment, Automated Qualitative Analysis

1 Introduction

Integrating Artificial Intelligence (AI), especially Large Language Models (LLMs), into safety-critical fields like aviation presents both transformative opportunities and significant challenges. Although LLMs can automate analysis and support decision-making, their ability to align with the nuanced, context-driven reasoning of human experts remains uncertain.

This study addresses this gap by comparing decision-making patterns among professional pilots, laypeople, and a local AI agent across hypothetical scenarios ranging from abstract dilemmas to realistic emergencies. Drawing on Prospect Theory and the Recognition-Primed Decision (RPD) model, we find that while AI can mimic human biases in abstract problems, it defaults to a risk-seeking, utilitarian logic that contrasts with aviation's safety-first, deontological principles.

Objectives: We (i) compare decision patterns and risk preferences across groups in abstract and operational dilemmas; (ii) analyse framing effects and expert intuition via Prospect Theory and RPD; (iii) identify misalignment points between AI logic and aviation safety principles; and (iv) propose design and alignment strategies (expert feedback, constitutional duty constraints, hard safety taboos, and continuous red-teaming).

Roadmap: Section 2 reviews Prospect Theory, the Recognition-Primed Decision (RPD) model, and the use of LLMs in safety-critical contexts. Section 3 details participants, scenarios, the procedural workflow (prompting, training, and LLM-as-a-Judge), and statistical methods; the ethics subsection closes Methods. Section 4 presents the results. Section 5 discusses expertise, ethics, implications, and practical utility. Section 6 consolidates limitations, risks, and mitigations. Subsection 6.3 outlines future work, and Section 7 concludes.

2 Background and Related Work

2.1 Prospect Theory and Framing

Prospect Theory explains systematic deviations from expected utility, including reference dependence (gains vs. losses), loss aversion, diminishing sensitivity, and probability weighting. In gain frames, people tend to be risk averse, whereas in loss frames they tend to be risk seeking; these effects replicate across domains and are sensitive to framing choices.[1, 2, 3, 4]

2.2 Expert Decision-Making (RPD) and Safety Culture

In high-risk, time-pressured environments such as cockpits, experts rely on recognition-primed decision-making: they rapidly recognise familiar patterns, generate a plausible course of action (often the first), and mentally simulate it rather than comparing multiple alternatives analytically.[5, 6, 7, 8] This sits within safety culture that encodes taboos around unacceptable risks and emphasises error management and resilience.[9, 10, 11, 12]

2.3 LLMs in Safety-Critical Domains / Dual-LLM

LLMs show strong pattern-matching on textual tasks but may hallucinate facts and over-optimise abstract objectives misaligned with domain norms. Dual-LLM pipelines (generator + judge) can improve output quality but may share biases if training data overlap, underscoring the need for domain constraints and human oversight.[13]

3 Methods

3.1 Participants and Groups

We analysed three decision-maker groups:

- **Pilots (Experts):** $N = 36$ professional pilots with flight experience ranging from 500–1000h to 1001–3000h, representing a benchmark for aviation decision-making shaped by training, experience, and internalised safety culture.[11, 12]
- **Lay Participants (Novices):** $N = 36$ participants with no flight experience (verified by the "Do you fly aircraft?" field), isolating domain-knowledge effects.
- **AI Agent:** A local LLM treated as a single-case study ($N = 1$), using a generator–judge pipeline (*Phi-3-Mini-Instruct* and *Zephyr-7B-Beta*). Conclusions specific to this system illustrate broader alignment challenges in critical domains.

3.2 Questionnaire and Scenarios (Q1–Q6)

Six key scenarios were posed:

- **Q1 (Operational risk):** Turning off the weather radar.
- **Q2 (Ethical dilemma):** Rescue trade-off (duty-of-care vs. utilitarian maximisation).
- **Q3 (Prospect Theory, loss frame):** Certain deaths vs. risky option.
- **Q4 (Operational risk):** Proceeding with a risk of serious failure.
- **Q5 (Operational risk):** Maintaining route through a storm with risk of structural damage.
- **Q6 (Prospect Theory, gain frame):** Certain saves vs. risky option.

Operational scenarios probe expert safety culture and taboo risks; PT scenarios isolate canonical framing effects; the ethical scenario tests duty-based vs. utilitarian priorities.

3.3 Procedure and Workflow

3.3.1 Prompting (Generator)

The generator (*Phi-3-Mini-Instruct*) produced structured outputs (MCQ, Likert, short text). Prompts specified incident context (single anonymised narrative), response formats, and constraints (avoid speculation; admit uncertainty). We used a low temperature (0.1) for consistency. Explicit instructions reduced, but did not eliminate, hallucinations.

3.3.2 Training (Generator and Judge)

Both *Phi-3-Mini-Instruct* (generator) and *Zephyr-7B-Beta* (judge) are Transformer-based models[14] fine-tuned locally without the specific incident narrative to avoid leakage. Hyperparameters are in Table 1. Corpora combined Wikipedia and aviation texts (sanitised for sensitivity), plus publicly available NTSB reports to improve domain familiarity. Validation used a calibration set for perplexity and macro-F1; early stopping controlled overfitting.

Table 1: Hyperparameters Used During Fine Tuning

	Phi-3-Mini-Instruct	*Zephyr-7B-Beta*
Model Size	3B parameters	7B parameters
Batch Size	8	4
Learning Rate	2×10^{-5}	1×10^{-5}
Epochs	3	2

3.3.3 LLM-as-a-Judge and Pseudocode

The judge LLM independently rated generator answers for *confidence*, *completeness*, and *groundedness* (0–1). To mitigate shared biases, we spot-checked 40 sampled items with human experts.

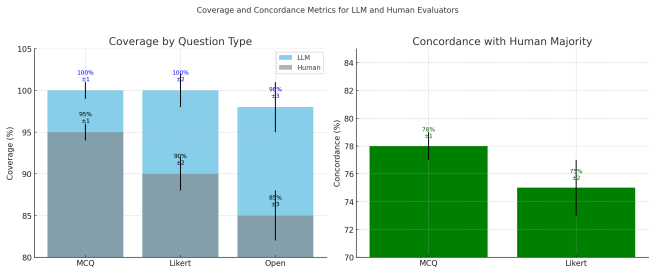

Figure 1: Local LLM-based pipeline with a judge LLM for partial self-evaluation.

Algorithm 1: Load Reports and Parse Questions

Input: *reportsDir*, *questionsFile*
Output: **reports**, **questions**

```
parser ← DataParser();
reports ← parser.loadReport(reportsDir);
questions ← parser.parseQuestions(questionsFile);
return reports, questions;
```

This routine initialises the data layer and returns two canonical objects consumed by the downstream pipeline.

- *Inputs.* `reportsDir` points to a directory with one or more anonymised incident narratives (pre-converted to text where necessary). `questionsFile` is a machine-readable definition of the questionnaire covering Q1–Q6 (e.g., JSON/CSV), including item IDs, item types (MCQ/Likert/open), and Safe vs. Risk mappings.
- *Steps.*
  1. `DataParser()` constructs a parser with normalisation policies (encoding, whitespace, tokenisation, redaction rules).
  2. `parser.loadReport(reportsDir)` ingests and cleans the narratives, applies basic quality checks (e.g., deduplication, language/encoding normalisation), and returns a structured container (list or dict) keyed by report ID.
  3. `parser.parseQuestions(questionsFile)` reads the questionnaire specification and builds a typed schema for each item: `id` $\in$ {Q1,...,Q6}, `type`, `options`, `safe/risk` labels, and any scoring rubrics.
- *Outputs.* **reports** is a cleaned corpus ready for prompting; **questions** is a validated catalogue of Q1–Q6 with consistent IDs, types, and option semantics.
- *Invariants and checks.* Guarantees a one-to-one mapping to {Q1,...,Q6}; asserts the presence of Safe/Risk tags where applicable; logs any malformed entries for audit; preserves anonymisation.

Algorithm 2: LLM-Based Analysis Pipeline

Input: **reportText**, **questions**

```
handler ← LocalModelHandler();
orchestrator ← LLMOrchestrator(handler);
foreach question ∈ questions do
    strategy ← StrategyFactory.getStrategy(question,
      reportText);
    prompt ← strategy.generatePrompt();
    rawAnswer ← handler.runGeneration(prompt);
    parsedAnswer ← strategy.parseResponse(rawAnswer);
    evalPrompt ← buildEvalPrompt(parsedAnswer, question,
      reportText);
    metrics ← handler.runEvaluation(evalPrompt);
    storeResult(question.id, parsedAnswer, metrics);
```

This loop orchestrates generation and judging for each questionnaire item, producing structured answers and quality metrics.

- *Initialisation.* `LocalModelHandler()` encapsulates the local generator and judge models (I/O, seeds, temperature, safety constraints). `LLMOrchestrator(handler)` coordinates per-item strategies and logging.
- *Per-question cycle.*
  1. `StrategyFactory.getStrategy(...)` selects a handling policy based on item type and context (e.g., MCQ vs. Likert vs. open), including how to inject `reportText`.
  2. `strategy.generatePrompt()` assembles a prompt with the required format, context window, and guardrails (e.g., "avoid speculation; admit insufficient information"), ensuring consistent phrasing across Q1–Q6.
  3. `handler.runGeneration(prompt)` queries the generator (low temperature for stability) and returns `rawAnswer`.
  4. `strategy.parseResponse(rawAnswer)` normalises the output into a canonical schema (e.g., MCQ option ID, Likert integer, free-text rationale), enforcing validity (only allowed options, required fields present).
  5. `buildEvalPrompt(parsedAnswer,...)` creates a focused evaluation prompt for the judge, providing the answer, the original question spec, and salient snippets from `reportText`.
  6. `handler.runEvaluation(evalPrompt)` obtains judge scores (e.g., confidence, completeness, groundedness in [0,1]) and any flags (contradictions, missing evidence).
  7. `storeResult(question.id,...)` persists the structured answer and metrics with metadata (timestamps, model versions, seeds) for auditability and later statistical analysis.
- *Reproducibility & safety.* Deterministic seeds and fixed templates improve reproducibility; validation catches out-of-schema outputs; logs enable traceability; the design allows swapping models/strategies without altering the data contract.

Multimodal data (scope and challenges). Typical investigations integrate multiple sources (CVR, FDR, meteorology). Audio–text and data fusion can reduce hallucinations but face privacy (CVR sensitivity), synchronisation (CVR/FDR timing), and access constraints (restricted raw data).[15] Referencing flight parameters or weather can prevent unfounded inferences (e.g., "engine fire" without corroborating data).

3.4 Statistical Analysis

The comparison of risky-choice proportions across groups was carried out using Fisher's exact tests, with Holm–Bonferroni correction across the six scenarios. Effect sizes are reported as Cramer's V with 95% confidence intervals. Likert-scale comparisons used Mann–Whitney (or Kruskal–Wallis for $k > 2$), with effect size r and bootstrap confidence intervals (10,000 resamples). Given the AI agent is a single-case study (N=1), we report its outcomes descriptively and avoid inferential statistics on that unit.[16, 17, 18, 19, 20]

3.5 Ethical and Regulatory Considerations

This study did not require Institutional Review Board approval because no sensitive or personally identifiable information was collected or processed. All data were fully anonymised; participation was voluntary with informed consent. All LLMs were run locally with no cloud dependencies, and no PII is stored in model weights. Although the pipeline shows time-saving potential, the single-incident scope means outputs are advisory only; certified investigators retain final authority. Deployment at scale must comply with ICAO/EASA/FAA procedures.

4 Results

Figure 2 summarises the proportion of risky choices across scenarios (Q1–Q6) by group.

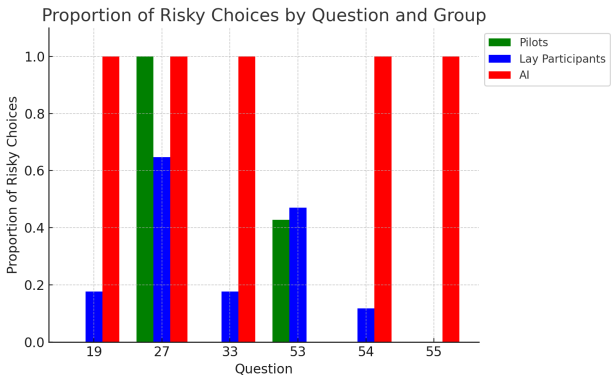

Figure 2: Proportion of risky choices by group (Pilots: green; Lay Participants: blue; AI: red) across Q1–Q6.

Pilots showed high consistency: marked risk aversion in **operational** scenarios (Q1, Q4, Q5) and framing-consistent behaviour in **Prospect Theory** scenarios (Q3 loss frame, Q6 gain frame). Lay participants displayed intermediate patterns with noticeable framing effects. The AI agent selected the risky option in all three operational scenarios (Q1, Q4, Q5) and followed canonical framing in Q3 (risk seeking) and Q6 (risk aversion).

4.1 Scenario-by-Scenario Group Comparisons

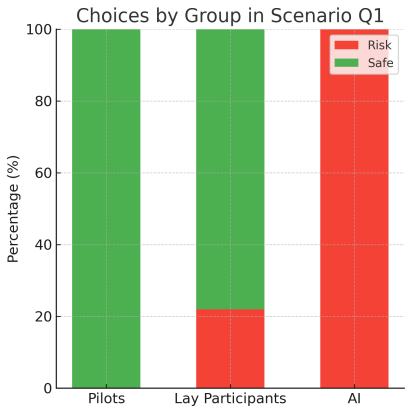

Figure 3: Q1 (Weather radar): pilots overwhelmingly chose the safe option; the AI selected the risky option.

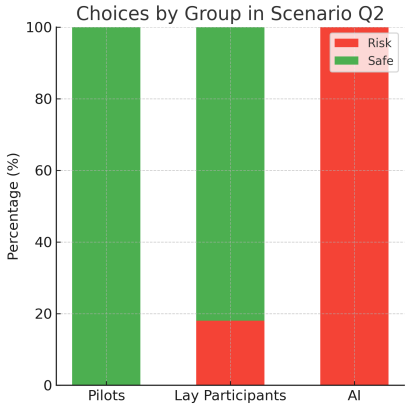

Figure 4: Q2 (Ethical rescue): pilots and most lay participants chose the safe option; the AI selected the risky alternative (aggregate utility).

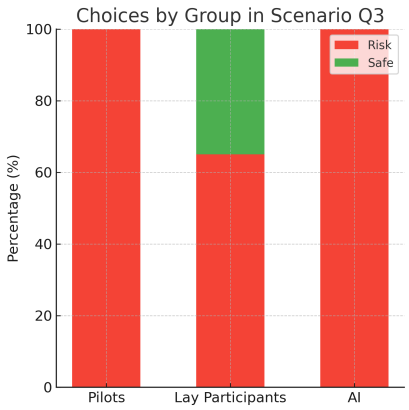

Figure 5: Q3 (Loss frame): strong risk-seeking across groups, consistent with Prospect Theory.

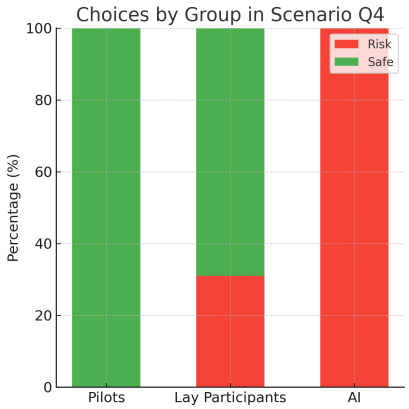

Figure 6: Q4 (Serious failure risk): pilots chose the safe option; the AI remained risk-seeking.

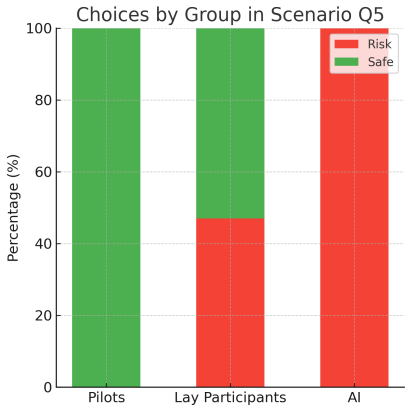

Figure 7: Q5 (Storm/structural risk): pilots avoided risk; the AI was risk-seeking.

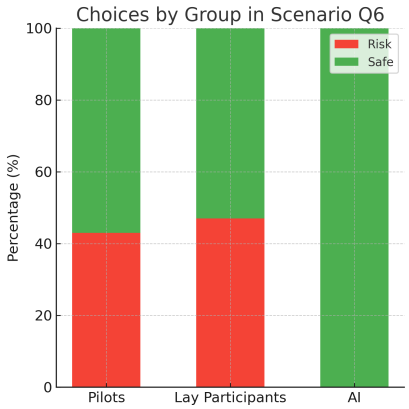

Figure 8: Q6 (Gain frame): all groups showed increased risk aversion, with the AI fully risk-averse.

Summary Table. Table 2 reports the percentage of risk-seeking choices for each group across Q1–Q6 (text and table values are consistent).

Table 2: Risk-seeking rates (%) by group for Q1–Q6

	Q1	Q2	Q3	Q4	Q5	Q6
Pilots	0	0	100	0	0	43
Lay Participants	22	18	65	31	47	47
AI Agent	100	100	100	100	100	0

Statistical Notes. Across operational scenarios (Q1, Q4, Q5), pilot vs. non-pilot and pilot vs. AI risky-choice proportions differed significantly by Fisher's exact tests with Holm–Bonferroni correction (all adjusted $p < .005$); effect sizes were moderate-to-large (Cramer's V). For the ethical dilemma (Q2), pilots vs. AI also differed significantly after correction (adjusted $p < .01$). Likert-scale contrasts (where applicable) yielded consistent Mann–Whitney results (effect size r) with 95% bootstrap CIs. The AI (N=1) is summarized descriptively without inferential testing.

5 Discussion

5.1 Prospect Theory vs. Operational Scenarios

In Q3 (loss frame) and Q6 (gain frame), humans and the AI exhibited canonical framing effects, consistent with Prospect Theory.[1, 2, 3, 4] However, in operational scenarios (Q1, Q4, Q5), pilots uniformly chose safe options, consistent with RPD-style pattern recognition and safety taboos that categorically proscribe certain hazards.[5, 6, 9, 10] Crew accounts in the real incident motivating the study illustrate rapid recognition and immediate safe action ("I saw the EGT spike... it's real"; "declared MAYDAY immediately"), consistent with recognition-primed scripts rather than analytical trade-offs.[21]

5.2 Ethical Tensions (Duty of Care vs. Utilitarianism)

In Q2, pilots and most lay participants chose to preserve lives under their protection (duty-of-care), whereas the AI selected a utilitarian maximisation. This aligns with consequentialist reasoning patterns observed in LLMs,[22, 23, 24, 25] but conflicts with professional deontological imperatives in aviation and rescue.[26, 27, 28] The divergence exemplifies value misalignment: the AI optimises expected outcomes while neglecting role-based duties and the ethics of acts vs. omissions.[29, 30, 31, 32]

5.3 Implications for Alignment

The generator–judge setup improves surface quality but risks shared biases if models share pretraining/fine-tuning distributions. Mitigations include using distinct model families, multiple independent judges, and human spot checks. More fundamentally, misalignment arises from (i) training data underrepresenting taboo risk patterns, (ii) context-intake that treats hazards as abstract probabilities rather than unacceptable categories, and (iii) prompting lacking explicit duty constraints. These require domain-specific alignment that encodes safety culture and duties, not only general helpfulness.

5.4 Practical Utility and Innovations

Even at this single-incident stage, the framework offers: (1) time-saving on routine coding; (2) consistency across recurring categories; (3) preliminary hypothesis checks by contrasting LLM "first pass" with investigator judgement; (4) a path toward adaptive or conversational PT processes; and (5) potential for multimodal integration (CVR/FDR/maintenance logs) to minimise speculation.

Linking model failures to training, context intake, and prompting. Finally, we explicitly trace the observed divergences to three loci: (i) **training data**, which underrepresent taboo-risk patterns and role-based duties encoded in aviation culture; (ii) **context intake**, which frames hazards as graded probabilities rather than categorical no-go conditions; and (iii) **prompting**, which omits explicit duty constraints and acceptable-risk thresholds. Together, these factors explain why the model reproduces canonical framing effects in abstract problems yet fails under operational, duty-bound scenarios. Our mitigations—expert-guided RLHF, constitutional duty constraints, hard safety taboos, and domain red-teaming—are designed to act on these loci (see Section 6).

6 Limitations, Risks and Mitigations

6.1 Risks of LLM Hallucination / Misalignment (and Study Limitations)

Observed error types: (1) factual hallucinations (systems not on the aircraft), (2) misattribution of roles (captain vs. first officer), (3) over-generalisations (e.g., "all checklists completed" where only partial completion was implied). With a single narrative, the model may extrapolate missing details; aircraft-specific idiosyncrasies are easily missed.
Study scope limitations: single incident; fixed 6-item questionnaire (no adaptive follow-ups); limited external validity across aircraft, weather, phases, and organisational cultures. These instrument and sampling limits constrain generalisation.

6.2 Mitigations (linking to training, context-intake, and prompting)

RLHF with experts (RLHF-E). Encode RPD-style heuristics and safety culture using pilot/engineer feedback, scoring compliance with taboo patterns rather than generic helpfulness.[33, 34, 5, 35]

Constitutional AI with duty constraints. Incorporate explicit rule sets (FAA *Pilot's Handbook of Aeronautical Knowledge*, QRH, regulations) to constrain actions that violate duties regardless of abstract utility; include role-based duties and "acts vs. omissions" ethics.[36, 37, 38, 39]

Hard constraints and formal verification. Translate non-negotiable safety taboos (e.g., "never intentionally degrade critical flight systems") into verifiable specifications (e.g., temporal/deontic logic) enforced architecturally or via wrappers.[40, 41, 42, 43]

Domain-specific red-teaming. Continuous adversarial testing by aviation experts (not only AI researchers) to surface failure modes absent from training, with an iterate-and-fix loop.[25, 44, 45]

Prompting and context-intake adjustments. Reinforce instructions to declare uncertainty; integrate cross-checks against known aircraft specs and constraints; train with negative examples where "insufficient info" is correct.

6.3 Future Work

To overcome the constraints of this pilot, we aim to:

- **Scale to Multiple Incidents**: We intend to analyze 50+ diverse incidents, including different aircraft classes, flight phases, and severities, and incorporate effect size reporting on a broader dataset.
- **Multi Modal Data**: Integrate cockpit voice recordings and flight data logs to reduce reliance on textual speculation.
- **Stronger Ethical Protocols**: Explore building an "ethical check" submodule or ensemble model that queries domain experts about morally nuanced scenarios.
- **Hybrid Evaluation**: Combine multiple LLM judges with partial domain expert arbitration, mitigating self reinforcement biases.

These expansions will help refine the pipeline so it can handle a wide range of scenarios with robust reliability and regulatory acceptance.

6.4 Multi Incident Plans

To strengthen external validity and increase inference power, our upcoming efforts will:

1. **Incorporate multiple incidents** (potentially 10 to 50 or more), covering a broad range of aircraft, phases of flight, and severities.
2. **Examine geographical diversity** (incidents from multiple airlines or regions), testing whether the LLM can adapt to distinct operational cultures.
3. **Employ temporal stratification** (incidents from different time periods), verifying how evolving standard operating procedures or regulatory updates may affect performance.

Such expansions would permit more robust statistical analyses using appropriate non parametric tests and effect size measures, thereby facilitating more reliable conclusions.

7 Conclusion

This study demonstrates that, although AI can mimic human biases in abstract risk scenarios, its alignment collapses in real-world operational contexts. Expert pilots consistently rejected risky options, guided by deeply internalised safety culture and

rapid pattern recognition, whereas the AI defaulted to risk-seeking, utilitarian logic. In ethical dilemmas, AI's maximising approach clashed with the deontological duty of care prioritised by humans. The main barrier to safe AI integration in critical domains is not computational capacity, but the absence of tacit cultural and ethical knowledge. Effective alignment will require codifying domain-specific values and continuous expert oversight. While the multi-layered alignment strategies proposed here are promising, further research covering more diverse incidents, AI models, and expert groups is essential before AI can be safely entrusted with safety-critical decisions in aviation or similar fields.

A Questionnaire and Coding Map (Q1–Q6)

Q1 (Operational). *Turning off weather radar.* Safe = keep radar on / avoid degrading situational awareness; Risk = turn off radar.
Q2 (Ethical). *Rescue trade-off.* Safe = return with 20 already rescued; Risk = attempt to save 30 more with 50% chance of losing all.
Q3 (PT Loss). *Certain deaths vs. risky option.* Safe = accept certain loss; Risk = gamble to reduce losses.
Q4 (Operational). *Proceed with serious failure risk.* Safe = do not proceed / mitigate; Risk = proceed.
Q5 (Operational). *Maintain route through storm (structural risk).* Safe = avoid storm; Risk = continue through storm.
Q6 (PT Gain). *Certain saves vs. risky option.* Safe = accept certain saves; Risk = gamble for higher saves.

B Generator Prompt and Judge Template (Summary)

Generator (Phi-3-Mini-Instruct). Include: incident context; response format (MCQ/Likert/open); constraints (avoid speculation, admit uncertainty); temperature=0.1.
Judge (Zephyr-7B-Beta). Score confidence, completeness, groundedness (0–1); require justification snippets; flag contradictions or missing evidence.

C Aggregated Data (Anonymized)

Per venue policy, aggregate tables and code to reproduce figures can be provided to reviewers upon request. All personally identifiable information is removed.

References

[1] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

[2] Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992.

[3] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[4] Nicholas C. Barberis. Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, 27(1):173–96, 2013.

[5] Gary A. Klein. A recognition-primed decision (rpd) model of rapid decision making. *Decision making in action: Models and methods*, 5(4):138–147, 1993.

[6] Gary Klein. The recognition-primed decision (rpd) model: Looking back, looking forward. *Naturalistic Decision Making*, 5(2):103–124, 1997.

[7] Gary Klein. Naturalistic decision making. *Human factors*, 50(3):456–460, 2008.

[8] Judith Orasanu and Terry Connolly. The reinvention of decision making. *Decision Making in Action: Models and Methods*, pages 3–20, 1993.

[9] James Reason. *Human Error*. Cambridge University Press, 1990.

[10] Robert L. Helmreich. Culture, error and crew resource management. *Safety in aviation: The management commitment*, pages 71–91, 2000.

[11] James Reason. *Managing the Risks of Organizational Accidents*. Ashgate, 1997.

[12] Sidney Dekker. *Drift into Failure: From Hunting Broken Components to Understanding Complex Systems*. CRC Press, 2011.

[13] Sébastien Bubeck et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[15] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021.

[16] S.S. Shapiro and M.B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[17] M. Hollander, D.A. Wolfe, and E. Chicken. *Nonparametric Statistical Methods*. Wiley, 3rd edition, 2013.

[18] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2nd edition, 1988.

[19] D.C. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, 10th edition, 2020.

[20] A. Field. *Discovering Statistics Using SPSS*. SAGE Publications, 3rd edition, 2009.

[21] Anônimo. Relato do comandante - incidente aéreo fictício para pesquisa, 2024. Documento fornecido para a análise.

[22] John Stuart Mill. *Utilitarianism*. Parker, Son, and Bourn, 1863.

[23] Jeremy Bentham. *An Introduction to the Principles of Morals and Legislation*. T. Payne and Son, 1789.

[24] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

[25] Ethan Perez et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

[26] Immanuel Kant. *Groundwork of the Metaphysics of Morals*. 1785.

[27] William David Ross. *The Right and the Good*. Clarendon Press, 1930.

[28] Tom L. Beauchamp and James F. Childress. *Principles of Biomedical Ethics*. Oxford University Press, 8th edition, 2019.

[29] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[30] Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

[31] Alan Gewirth. Professional ethics: The separatist thesis. *Ethics*, 96(2):282–300, 1986.

[32] Bernard Gert. *Common Morality: Deciding What to Do*. Oxford University Press, 2004.

[33] Long Ouyang et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[34] Daniel M Ziegler et al. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

[35] Dario Amodei et al. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[36] Yuntao Bai et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[37] Amanda Askell et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

[38] Federal Aviation Administration. *Pilot's Handbook of Aeronautical Knowledge*. U.S. Department of Transportation, 2016. FAA-H-8083-25B.

[39] Amelia Glaese et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

[40] Guy Katz et al. Reluplex: An efficient smt solver for verifying deep neural networks. *In International Conference on Computer Aided Verification*, pages 97–117, 2017.

[41] Edmund M. Clarke, Orna Grumberg, and Doron A. Peled. *Model Checking*. The MIT Press, 1999.

[42] Dan Hendrycks et al. What would jiminy cricket do? towards a theory of moral machine action. *arXiv preprint arXiv:2110.13452*, 2021.

[43] Anders Sandberg. Ethics of artificial intelligence. *The Cambridge Handbook of the Ethics of Human-Robot Interaction*, pages 58–76, 2018.

[44] Deep Ganguli et al. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

[45] Nicholas Carlini and David Wagner. Adversarial examples are not bugs, they are features. *2019 IEEE Security and Privacy Workshops (SPW)*, pages 12–18, 2019.