RISK JUSTIFICATION REPORT

AI Decision Abstention & Escalation Framework

1. Introduction

Artificial Intelligence systems deployed in high-impact domains must balance automation efficiency with safety, ethics, and reliability. The most severe AI failures often occur not due to incorrect predictions alone, but because the AI acted when it should have abstained.

This report justifies the design of an AI Abstention & Escalation Framework for a Healthcare Symptom Triage Assistant, focusing on preventing irreversible harm and ensuring responsible decision-making.

2. System Overview

- **System Type:** Assistive Healthcare AI
- **Function:** Provide preliminary urgency guidance based on symptoms
- **Role of AI:** Decision-support only (not diagnostic authority)

Possible AI Actions:

- Provide recommendation
- Abstain from decision
- Escalate to human oversight

3. Identified Risks

3.1 Safety Risks

Risk Category	Description
False reassurance	AI classifies severe condition as non-urgent.
Missed emergencies	Low-confidence but high-risk symptoms missed by automation.
Overconfidence	AI acts despite significant uncertainty in the model output.
Data ambiguity	Incomplete or conflicting symptoms lead to hazardous outputs.

3.2 Ethical Risks

- **Automation bias:** Users may trust the AI blindly, overriding personal caution.
- **Replacement of clinical judgment:** Risk of over-reliance on silicon over professional expertise.
- **Vulnerable populations:** Unequal impact or bias in diagnostic accuracy for specific demographics.

3.3 Legal & Compliance Risks

- Liability for unsafe or incorrect medical recommendations.
- Regulatory violations (e.g., FDA, MHRA, or local healthcare software standards).
- Failure of duty of care to the end-user.

3.4 Trust & Business Risks

- Loss of user confidence following a public error.
- Long-term brand damage to the healthcare provider or developer.
- Reduced global adoption of AI-driven tools due to safety fears.

4. Justification for Abstention

4.1 Why Abstention is Necessary

Abstention prevents the AI from making decisions when confidence is low, data quality is insufficient, harm potential is high, or the situation exceeds the model's trained knowledge base.

> **Core Principle:** Not acting is safer than acting incorrectly in high-risk contexts.

4.2 Risk Reduction Benefits

Scenario	If AI Acts	If AI Abstains
Low confidence prediction	Incorrect advice/recommendation	Route for Human Review
Conflicting symptoms	Misclassification of severity	Expert Judgment required
Rare critical condition	False reassurance	Immediate Escalation

5. Justification for Escalation

Escalation is required when harm severity is High or Critical, when uncertainty intersects with serious outcomes, or when ethical responsibility demands human judgment.

Key Rationale: Humans handle contextual nuance more effectively, accountability remains clear, and the likelihood of catastrophic failure is significantly reduced.

6. Confidence vs Harm Trade-Off

6.1 Automation Limitation

High confidence does NOT guarantee safety. For example, a 92% confidence rating coupled with a Critical harm potential is still considered too unsafe to automate without oversight.

6.2 Safety Rule

Condition	Decision
Low confidence	Abstain
High harm severity	Escalate
Low harm + high confidence	AI Acts

7. Speed vs Safety Trade-Off

Factor	Full Automation	With Abstention Framework
Response Speed	Very Fast	Slight Delay
Safety	Lower	Higher
Risk Exposure	Higher	Reduced
Trust	Fragile	Stronger

Conclusion: Minor delays are justified to prevent irreversible harm.

8. Ethical Justification

The framework aligns with the core principles of Bioethics:

- **Beneficence:** Active effort to provide benefit and avoid harm.
- **Non-maleficence:** Preventing unsafe actions.
- **Human Oversight:** Ensuring humans remain the ultimate authority in medical care.

Abstention reflects ethical humility — the system explicitly acknowledging its own limitations.

9. Accountability & Responsibility

Event	Responsible Party
AI automated decision	AI System & Organization
AI abstains	Organization (via design choice)
Escalated decision	Human Reviewer
Systemic failure	AI Governance / Safety Team

10. Business Justification

Although abstention may reduce the raw automation rate and slightly increase operational costs (due to human-in-the-loop requirements), it provides significant value by:

- Reducing legal and liability risks.
- Improving user trust and retention.
- Preventing catastrophic failure-related news.

- Supporting rigorous regulatory compliance.

11. Design Assumptions

- AI is assistive, not diagnostic.
- Confidence scores are mathematically calibrated for reliability.
- Harm severity categories are strictly defined by clinical standards.
- Human oversight is available and accessible.
- No real patient data was used in the formulation of this conceptual framework.

12. Final Conclusion

A responsible AI system must know: **When to act, When to abstain, and When to escalate.**

This framework prioritizes safety over blind automation, ensuring that the deployment of the Healthcare Symptom Triage Assistant remains ethical, reliable, and trustworthy.