International Conference on Science, Engineering & Management Trends

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Explainable AI for Detecting Safety Barrier Degradation from Near Miss Reports

Dr. Ashish Mulraj Ved[1], Ravindra Kumar[2]

[1] PhD Research Scholar, Faculty of Commerce and Management, Asian International University, Manipur, India.
[2] Assistant Professor, Asian International University, Manipur, India.
1A HK International
Corresponding author: info@hkinternational.uk; drashishved@gmail.com

Abstract- Industrial maintenance generates near miss narratives that may reveal weaknesses in energy isolation, equipment integrity, permit coordination and other safety barriers. However, interpreting these reports requires contextual judgement, and an automated summary can conceal missing evidence or produce unjustified conclusions. This conceptual paper proposes an evidence-linked artificial intelligence framework for identifying reported barrier degradation and supporting accountable human review. The framework combines a controlled barrier taxonomy, source-linked extraction, retrieval from approved site documents, uncertainty handling and verification of corrective actions. Its central research proposition is that the value of AI should be evaluated through the quality of barrier decisions and the completeness of supporting evidence, rather than classification accuracy alone. A staged mixed methods protocol is specified, including expert annotation, temporal and site-separated testing, comparison with simpler methods, a randomised reviewer study and a prospective shadow deployment. Primary outcomes concern detection of expert-adjudicated barrier concerns, unsupported assertions and reviewer decision quality. Incident reduction is reserved for later investigation and is not inferred from report processing performance. No field dataset, implemented prototype or empirical results are presented. The contribution is a testable research design that connects explainable AI with practical safety assurance while retaining operational authority with competent personnel

Keywords- occupational safety; process safety; near miss analysis; safety barriers; explainable artificial intelligence; retrieval augmented generation; human oversight; industrial maintenance

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

I. INTRODUCTION

Near miss reports can preserve observations that are difficult to capture in a standard inspection score. A maintenance worker may describe pressure remaining after isolation, a missing permit attachment, or an alarm that was ignored because it had activated repeatedly. Such accounts can help direct investigation, but they do not establish causes by themselves. The UK Health and Safety Executive investigation workbook structures learning around information gathering, analysis, identification of controls and implementation of an action plan [1]. The proposed study addresses the information analysis step and its connection to verified action.

The research setting is contractor-supported mechanical maintenance at industrial facilities. The initial scope covers energy isolation, pressure systems and permit interfaces. These activities provide a bounded environment in which barrier requirements can be linked to identifiable tasks and approved local documents. Broad claims about all HSE activities would be premature; occupational disease surveillance, environmental impact modelling and autonomous process control fall outside this paper.
The research question is whether an evidence-linked AI assistant can help competent reviewers recognise reported degradation of safety barriers without introducing unsupported explanations or excessive review workload. A safety barrier is understood here as a technical, organisational or human measure intended to prevent an unwanted event or limit its consequences. Barrier degradation means a reported or independently verified reduction in that measure's availability, effectiveness or implementation. A report suggesting degradation is a reason to investigate, rather than proof that a barrier has failed.

II. Conceptual foundations and research gap

1. Barrier assurance and incident learning

HSE guidance on process safety indicators links performance measurement to the operation of risk control systems [2]. This provides a useful basis for moving beyond counts of reports toward questions about particular controls. In this paper, the proposed extension is to associate narrative evidence with a named barrier, its local requirement, its responsible owner and its verification record. The association is a design proposal; HSG254 does not validate the AI system described here.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

The hierarchy of controls provides a separate check on proposed responses. NIOSH identifies elimination, substitution, engineering controls, administrative controls and personal protective equipment as a preferred sequence [3]. A reviewer should therefore be able to challenge an AI suggestion that merely repeats training or PPE advice where a higher-order control warrants investigation. The assistant should retrieve relevant options, while a competent person determines technical suitability and any new risks.

2. AI grounding and explanation

Retrieval augmented generation combines a language model with access to an external information collection. Lewis and colleagues demonstrated this approach for knowledge-intensive language tasks [4]. That evidence motivates investigation of document-grounded assistance, but it does not establish effectiveness for industrial safety. Retrieval can return an irrelevant, obsolete or inapplicable passage; generation can also misinterpret an appropriate passage. Both stages require separate assessment.

Ribeiro and colleagues introduced LIME to explain individual classifier predictions through local interpretable approximations [5]. Their work illustrates an established approach to interrogating model behaviour. In the present framework, highlighted source passages and feature explanations serve different purposes. Source passages support factual traceability; a feature explanation describes model sensitivity. Neither establishes a causal account of an incident, and a fluent generated rationale should not be accepted as evidence of either.

NIST AI RMF 1.0 provides a lifecycle structure for managing AI risks [6], with a complementary profile addressing generative AI [7]. The proposed design draws on this governance perspective while specifying an application-level test: can a reviewer inspect the evidence, identify uncertainty, reject a suggestion and preserve responsibility for the decision? Adoption of a framework alone does not demonstrate that these capabilities work.

3.Proposed research contribution

This paper synthesises selected primary research and official guidance; it is not a systematic review. Consequently, it does not claim that barrier-oriented near miss analysis is unprecedented. The proposed contribution is an integrated evidence schema, abstention policy and evaluation protocol connecting narrative interpretation to independently verifiable barrier decisions. A doctoral literature review should

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

establish how this combination differs from existing incident classification, knowledge graph, barrier management and human factors research before asserting novelty.

4.Research objectives and propositions

The principal objective is to develop and evaluate an AI-assisted review process that identifies barrier concerns with traceable evidence. The unit of analysis is a reported event together with the barrier decisions made during its review. Multiple reports concerning one event must be linked so that repeated descriptions are not treated as independent evidence.

RQ1 asks how accurately the framework extracts barrier concerns and distinguishes reported facts from inference. RQ2 asks whether document grounding reduces unsupported assertions compared with the same model without retrieval. RQ3 asks whether assistance improves reviewer detection of important barrier concerns while preserving appropriate rejection of incorrect advice. RQ4 asks how performance and review burden vary across sites, report quality and contractor language patterns.

The corresponding propositions are testable rather than established findings. P1 is that grounding will reduce the proportion of unsupported factual assertions. P2 is that explicit barrier mapping will improve detection of adjudicated barrier concerns relative to a generic hazard classifier. P3 is that evidence displays with uncertainty cues will improve reviewer decision quality relative to unassisted review. P4 is that performance will deteriorate when local procedures or reporting language differ from the development setting, requiring external validation and recalibration.

IV. Proposed framework

1. Inputs and controlled knowledge

The input record comprises a pseudonymous event identifier, narrative, task, equipment category, report date, site and available attachments. Names and unnecessary identifiers should be removed before model processing. The document collection should contain approved and version-controlled procedures, permit requirements, isolation standards and equipment instructions that the organisation is permitted to use. Each passage must retain its document identifier, version, effective date, location and applicability limits.

A site-approved taxonomy should distinguish the hazard, initiating condition, preventive barrier, mitigating barrier, barrier status and evidence type. It should allow multiple barriers per report and an

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

explicit unknown state. Missing mention of an isolation check must be coded as unreported unless independent evidence establishes that the check was omitted. Taxonomy development should involve maintenance personnel, contractors and HSE specialists so that categories reflect operational language.

2. Analysis and evidence output

Processing should first identify and link duplicates, then extract candidate factual spans without adding explanations. A classifier proposes hazard and barrier labels. Retrieval locates applicable requirements, and an evidence checker tests whether each factual claim is supported by the narrative or a cited document. A proposed concern may remain unresolved if a procedure is unavailable, applicability is disputed or the report lacks necessary detail.

The reviewer receives a compact evidence record containing the original report excerpt, candidate barrier, retrieved requirement, interpretation, uncertainty and questions for verification. Each interpretation should be tagged as explicit report content, documentary requirement, reviewer-confirmed information or model inference. The interface should preserve contradictions rather than merge them into a confident summary. The approved output is a signed review record, not an automatically authorised work instruction.

Untrusted report text and attachments must be treated as data, including any embedded instruction to ignore rules or disclose records. Retrieval should be restricted by site and user permissions. Logs should record model version, prompt configuration, retrieved passages, reviewer amendments and subsequent verification. A model or document update should trigger targeted regression checks before reuse.

3. Abstention and operational authority

The assistant should abstain from recommending a barrier status when evidence is insufficient or conflicting, while routing the case for review. Abstention must not suppress an urgent concern: an existing site rule for suspected hazardous energy or a reported critical barrier failure should route the case immediately through established reporting arrangements. The research assistant must never issue permits, confirm zero energy, approve restart or change plant control settings.

Confidence should be estimated and assessed against held-out annotated cases. A model's self-reported confidence and retrieval similarity are not substitutes for calibrated error estimates. Routing thresholds should be chosen on validation data against a predefined missed-concern tolerance and an

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

achievable review workload, then frozen before testing. The safety owner should approve the operating policy and retain the ability to suspend the assistant.

4. Corrective action and verification

A competent reviewer decides what further evidence is required and whether action is necessary. An action record should identify the barrier, owner, deadline, acceptance criterion and verifier. Closing an action requires evidence that the agreed criterion was met; uploading a photograph or marking a task complete is not automatically sufficient. Where practical, verification should be performed by someone other than the person completing the action. Repeated concerns should prompt review of system conditions rather than automatic escalation against individual workers.

5. Illustrative application

The following example is hypothetical and does not describe a field observation. A report states: "After the pump was isolated, pressure was still indicated on the discharge gauge. Maintenance stopped and informed the supervisor. No injury occurred." The assistant should preserve those facts and propose an energy-isolation barrier concern for assessment. It should not assert that the valve leaked, that the gauge was accurate or that a worker failed to follow a procedure.

If an applicable site procedure requires verification of isolation before opening equipment, the assistant may retrieve that requirement and ask for the isolation record and verification evidence. If no approved procedure is available, it should disclose the gap. The reviewer then determines the appropriate investigation and operational response through existing site arrangements. Successful analysis means identifying the evidence gap and concern correctly; it does not mean predicting the technical cause.

Evidence item	Permitted interpretation	Unsupported conclusion
Pressure remained indicated	A possible isolation or indication concern requires review	The isolation valve definitely failed
Work was stopped	The report describes a protective response	Every required control was effective
No injury occurred	No injury is reported for this event	The potential consequence was minor
Verification record absent	Verification status is unknown in available material	Verification was definitely omitted

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

V. Proposed research methodology

1. Study stages and access conditions

The study would follow a staged mixed methods design. Stage 1 develops the taxonomy and interviews intended users about review decisions. Stage 2 performs retrospective technical validation using authorised, de-identified records. Stage 3 tests reviewer performance in a controlled experiment. Stage 4 evaluates workflow feasibility in prospective shadow mode, in which normal safety arrangements continue and research outputs do not control operations. No participating organisation or dataset has been secured for this paper.

The scope should initially be limited to a small number of maintenance settings with sufficiently documented barriers. Final recruitment, record counts and study duration must follow an access assessment and a prespecified precision or power analysis. An arbitrary large sample would not compensate for a scarcity of important barrier concerns. The feasibility stage should estimate concern prevalence, annotation effort, clustering by site and event, and the minimum effect on reviewer decisions that would justify adoption.

2. Annotation and reference standard

At least two appropriately experienced reviewers should independently label each evaluation event using a written codebook while blinded to model output. Labels should cover factual spans, hazard categories, barrier identity, evidence sufficiency, urgency of review and whether further investigation is required. A third expert should adjudicate disagreements. The reference standard is an expert interpretation of available records, not direct knowledge of all underlying conditions.

Before adjudication, agreement should be reported for categorical labels, including agreement specific to uncommon critical concerns. Raw agreement and a suitable chance-adjusted statistic should accompany the label distribution. Disagreements and ambiguous records should be retained in an uncertainty analysis rather than silently excluded. The assessment of operational verification, where available, should remain separate from the report-only annotation task.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

3. Splitting data and preventing leakage

Records describing the same event, duplicated text, related follow-up reports and repeated templates must remain within one split. Development should use earlier records; a later chronological period should be reserved for final testing. Where multiple sites are available, at least one site should be reserved for external evaluation. If this is impossible, claims must be limited to the development setting. At each test event, the retrieval collection should contain only documents available at that time. Investigation conclusions and later corrective-action records must not enter a model intended for initial triage. Preprocessing, prompts, classifier parameters, calibration and routing thresholds should be frozen before the final test. Any subsequent tuning requires a new untouched evaluation set.

4. Comparators and ablation tests

The minimum comparators are a documented keyword or rule system, a TF-IDF linear classifier, and the selected language model without retrieval. The proposed system adds controlled retrieval and evidence checks. A separate ablation should remove barrier mapping while keeping the underlying model and document access constant. This distinguishes the value of the application design from the value of a larger model or additional information.

All methods should receive equivalent input records and be scored against the same reference standard. Model identity, version, decoding settings, retrieval configuration and document collection date should be recorded. Repeated runs should quantify variability when generation is stochastic. Manual review is evaluated as a decision process in the reviewer experiment rather than treated as a single error-free benchmark.

5. Reviewer experiment

Competent reviewers should be randomised to assisted or unassisted review in a balanced crossover design using different but comparable case sets. Case order and condition order should be counterbalanced, and the same reviewer should not encounter the same event in both conditions. Independent assessors should score final decisions without knowing the review condition. Analysis should account for reviewer experience, case difficulty and repeated observations.

The primary human outcome is correct identification and escalation of adjudicated important barrier concerns. Secondary outcomes include unsupported decisions, time per case, appropriate rejection of erroneous suggestions and perceived workload. Deliberately seeded incorrect AI suggestions may be

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

used in a controlled simulation to measure overreliance, with consent and debriefing. Such suggestions must not be inserted into live operational advice.

VI. Evaluation and analysis plan

The protocol should prespecify one primary technical endpoint and one primary human endpoint to limit selective interpretation. The proposed technical endpoint is recall of expert-adjudicated important barrier concerns at a fixed, feasible review capacity. The human endpoint is the difference between conditions in the proportion of cases receiving the correct barrier-review decision. Both outcomes require explicit label and denominator definitions before data collection.

Measure	Definition	Interpretation constraint
Concern recall	Correctly flagged important concerns divided by all annotated important concerns	Report the count and uncertainty for rare concerns
Alert precision	Confirmed concern alerts divided by all concern alerts	Interpret jointly with review workload
Unsupported assertions	Unsupported factual assertions divided by all factual assertions in generated outputs	Independent assessors check source support
Abstention coverage	Cases receiving a substantive output divided by all cases	Show error rate alongside coverage
Verification completion	Actions meeting an independently checked criterion divided by actions due	A completed form is not proof of effectiveness

Technical results should include precision-recall curves, class-specific performance, false-negative case review and uncertainty intervals. Confidence calibration should be assessed only for outputs with a defined probabilistic target. Resampling should preserve event clusters and, where feasible, site structure. Paired comparisons are appropriate because competing systems process the same cases. Unsupported assertion scoring should distinguish absent support, contradiction and incorrect applicability of a cited requirement.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

The reviewer study should estimate effect sizes with confidence intervals using a model that accounts for reviewer and case variation. Time savings cannot compensate for a material increase in missed concerns. Before recruitment, the research team should define the acceptable safety margin and determine sample size for both decision quality and that constraint. Secondary and subgroup analyses should be identified as exploratory unless adequately powered.

Prospective shadow monitoring should assess latency, alerts per reviewer, unresolved cases, data access failures and changes in reporting behaviour. A decline in reported near misses could reflect reduced participation rather than improved safety. Injury frequency and major incidents are distal, uncommon outcomes; no causal reduction should be inferred from a short pilot or improved classification. A later intervention study would require a separate design, sufficient follow-up and adjustment for exposure and concurrent changes.

VII. Governance ethics and worker participation

Research involving worker interviews, identifiable records or simulated decision experiments should undergo the relevant institutional ethics review before recruitment or access. Data access agreements should specify purpose, confidentiality, retention, permitted model providers and whether records may be used for training. The organisation should assess re-identification risks because task, shift and equipment details may identify a worker even after names are removed.

Workers and contractor representatives should help design the reporting and review process. Individual productivity scoring and disciplinary ranking are outside the intended use. Participants should understand how information is used and how to challenge an interpretation. Research consent arrangements should account for employment power differences. Serious concerns encountered during the study should follow an agreed escalation process without waiting for research analysis.

Access controls, encryption, audit logs and separation of research data from operational personnel files should be specified before deployment. A change log should distinguish an updated model from an updated local procedure. Document expiry, unavailable retrieval, unexplained performance deterioration or repeated unsupported outputs should trigger review and possible suspension. These are proposed controls whose effectiveness must be tested, not assurances that the system is safe.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

VIII. Discussion and practical implications

The proposed framework changes the focus of near miss analytics from producing a category to supporting a defensible review decision. Its potential benefit lies in making recurring concerns and missing evidence easier to inspect. The test of usefulness is whether reviewers reach better-supported decisions under realistic workload constraints. A concise output with an explicit unknown may be more useful than a detailed explanation that invents a cause.

The design also makes several failure modes visible. Good classification performance can coexist with poor document applicability. Accurate source quotation can coexist with an incorrect interpretation. Reviewers may accept a polished recommendation despite contradictory evidence. Separating extraction, retrieval, interpretation and human decision evaluation allows each of these failures to be measured rather than hidden within a single accuracy figure.

An initial implementation could support an existing near miss review meeting rather than introduce a separate reporting platform. The organisation would need an approved document collection, a barrier owner, adequate reviewer capacity and a practical way to verify actions. Small organisations should compare the complete cost of annotation, document maintenance, secure processing and review against a simpler rule-based system. AI is justified only if the measured benefit exceeds this additional burden.

Transfer to oil drilling, construction or chemical processing would require new terminology, barrier requirements and external testing. Occupational exposure reports may also require specialist assessment and different time horizons. The present framework offers a bounded doctoral programme with linked outputs: a documented annotation scheme, reproducible benchmark, evaluated assistance interface and prospective feasibility assessment. Each output should be reported according to what was actually completed.

IX. Limitations

This is a conceptual paper and research protocol. It contains no collected dataset, implemented algorithm, empirical results or demonstrated operational improvement. Selected sources support its foundations but do not establish the novelty or effectiveness of the complete design. A systematic

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

literature review and feasibility work remain necessary. Report-based labels may reproduce reporting bias and cannot capture unreported hazards.Expert disagreement, sparse important events, differences in documentation and changes in work practices may limit inference. Site-specific procedures may restrict reproducibility, while public or constructed examples may not represent real reporting language. A successful laboratory evaluation would not establish safe live use. Even a prospective improvement in verified actions would require further evidence before attributing fewer incidents to the system.\

X.Conclusion

Evidence-linked AI analysis of near miss reports is a testable route for investigating how digital tools can support safety barrier assurance during industrial maintenance. This paper specifies a controlled workflow in which narrative facts, local requirements, uncertainty, human decisions and action verification remain distinguishable. Its proposed doctoral contribution lies in evaluating that connection rigorously. Claims of improved safety should follow independent testing of decision quality and operational outcomes; they cannot be inferred from the conceptual framework alone.

Declarations

- Research and data status: No primary dataset was generated or analysed for this paper. The study stages described are proposed future work. The illustrative case is hypothetical.
- Ethics: This manuscript reports no participant study. Institutional ethics review and site permissions should be obtained as applicable before the proposed empirical work begins.
- Authorship and disclosure: The named author should review and approve the technical content and confirm authorship responsibilities before submission. AI assistance was used in preparing this manuscript; the final disclosure should follow the chosen conference policy.
- Funding and interests: The author's listed affiliation is 1A HK International. Funding arrangements, relevant commercial interests and any further affiliations should be confirmed by the author before submission. No deployment, sponsorship or external endorsement is asserted.

References

1. Health and Safety Executive. Investigating accidents and incidents. HSG245. 2004. https://www.hse.gov.uk/pubns/books/hsg245.htm

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

2. Health and Safety Executive. Developing process safety indicators A step-by-step guide for chemical and major hazard industries. HSG254. 2006 https://www.hse.gov.uk/pubns/books/hsg254.htm
3. National Institute for Occupational Safety and Health. Hierarchy of Controls. 10 April 2024. https://www.cdc.gov/niosh/hierarchy-of-controls/about/index.html
4. Lewis P, Perez E, Piktus A, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems. 2020;33:9459–9474. https://arxiv.org/abs/2005.11401
5. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:1135–1144. https://arxiv.org/abs/1602.04938
6. Tabassi E. Artificial Intelligence Risk Management Framework AI RMF 1.0. NIST AI 100-1. 2023. https://doi.org/10.6028/NIST.AI.100-1Autio C, Schwartz R, Dunietz J, et al. Artificial Intelligence Risk Management Framework Generative Artificial Intelligence Profile. NIST AI 600-1. 2024. https://doi.org/10.6028/NIST.AI.600-1
7. Web sources accessed 19 October 2026.