Ariyan Azami

Lassonde School of Engineering, York University

4700 Keele Street, Toronto, Ontario M3J 1P3

July 22, 2026

Course Director, ENG 2003: Effective Engineering Communication

Lassonde School of Engineering, York University

Dear Course Director,

Please find enclosed the final technical report entitled *Algorithmic Fairness in Automated Decision Systems: Explainable AI as a Mitigation Strategy for Bias in Hiring and Credit-Scoring Pipelines*, prepared in fulfilment of Phase 4 (Revise, Finalize and Reflect) of Term Project 1 for the Summer 2026 term. This version revises the Phase 2 draft in response to feedback received from three peer reviewers through the PeerScholar platform.

The report examines a grand challenge of the twenty-first century identified in the BBC Future expert survey: the growing use of poorly understood, data-driven algorithms in high-stakes decisions such as hiring and lending. It proposes explainable artificial intelligence, and specifically the post-hoc attribution methods SHAP and LIME, as a practical bias-mitigation and audit strategy within the discipline of computer engineering. The report explains the mechanics of both methods, evaluates their strengths and limitations against current regulatory and legal developments including the EU AI Act and Mobley v. Workday, and argues that explainable AI is a necessary, though not sufficient, layer in fair automated decision systems.

A revision summary documenting the peer feedback and the changes made since the draft appears in Appendix A, and a disclosure of AI use appears in Appendix B. Thank you for your time and consideration; I welcome any further feedback.

Sincerely,

Ariyan Azami

Student ID: 222004071

Algorithmic Fairness in Automated Decision Systems

Explainable AI as a Mitigation Strategy for Bias in Hiring and Credit-Scoring Pipelines

Term Project 1: Final Technical Report (Phase 4)

ENG 2003: Effective Engineering Communication

Prepared by: Ariyan Azami (Student ID: 222004071)

Prepared for: ENG 2003 Instructional Team

Lassonde School of Engineering, York University

Discipline: Computer Engineering

Date: July 22, 2026

Executive Summary

Automated decision systems now mediate who is shortlisted for a job and who is approved for credit. Experts surveyed by BBC Future identify the opacity and unexamined fairness of these data-driven systems as a grand challenge of the twenty-first century: algorithms trained on historical data can silently reproduce and amplify discrimination at scale, while remaining inscrutable to the people they affect and the regulators who oversee them [1]. The consequences are no longer hypothetical. In Mobley v. Workday, a United States federal court allowed discrimination claims to proceed against an AI hiring-software vendor after a plaintiff was algorithmically rejected from over one hundred positions [6], and the European Union's AI Act now classifies both employment and credit-scoring systems as high-risk, imposing binding transparency and human-oversight obligations [5].

This report proposes explainable artificial intelligence, and specifically the two dominant post-hoc attribution methods, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), as a practical mitigation strategy within computer engineering and applied AI. These methods quantify how much each input feature contributed to an individual model decision, converting an opaque prediction into an auditable evidence trail. The report explains how each method works, where each is appropriate, and how attribution outputs enable three concrete interventions: detecting proxy discrimination during audits, generating the individualized explanations that regulators increasingly require, and supplying the documentary evidence needed in litigation and compliance reviews.

The analysis concludes that explainable AI is a necessary but not sufficient layer of a fair decision pipeline. SHAP and LIME operate after a model is trained; they expose bias but do not remove it, and their outputs can be unstable or misread by non-expert users. The report therefore recommends deploying explainable AI as a mandatory audit layer coupled with human oversight, standardized fairness metrics, and data-level interventions: a combined socio-technical approach that peer-reviewed syntheses in both hiring [2] and credit scoring [3] independently endorse.

Table of Contents

List of Figures

List of Tables

5. Introduction and Background

5.1 The Challenge: Opaque Algorithms in High-Stakes Decisions

Among the fifty grand challenges for the twenty-first century compiled by BBC Future from expert interviews, danah boyd, Principal Researcher at Microsoft Research and founder of the Data and Society Research Institute, identifies a distinctly modern problem: data-driven technologies are increasingly embedded in consequential decisions, yet regulators, civil society, and the public demand that these systems be “fair” and “ethical” without a shared, operational definition of what those words mean when translated into code [1]. boyd's framing falls under the article's artificial-intelligence theme, and it captures the gap this report addresses: the systems making decisions about people are becoming more accurate and less intelligible at the same time.

Two application domains make the stakes concrete. In hiring, employers use algorithmic tools to screen résumés, score recorded video interviews, and rank candidates; a 2024 peer-reviewed synthesis catalogues how bias can enter each stage of such pipelines, from unrepresentative training data through proxy features to deployment feedback loops [2]. The canonical cautionary tale is Amazon's experimental résumé-screening engine, abandoned in 2018 after engineers found it systematically downgraded résumés containing the word “women's,” having learned the male-dominated hiring patterns embedded in a decade of historical applications [10]. In consumer lending, machine-learning credit models outperform traditional scorecards on predictive accuracy, but a 2026 systematic review of 43 studies finds that performance, fairness, and explainability are almost always studied in isolation, leaving deployed systems accurate but unexamined [3].

5.2 Why the Problem Is Urgent Now

Three recent developments have converted this from an academic concern into an engineering obligation. First, litigation. In July 2024, the United States District Court for the Northern District of California ruled in Mobley v. Workday that an AI screening-software vendor could plausibly be treated as an agent of the

employers using its tools, allowing disparate-impact claims under Title VII, the ADEA, and the ADA to proceed [6]. The plaintiff, an African American man over the age of forty who also has a disability, alleged that he was rejected from more than one hundred positions, in some cases within an hour of submitting an application, a speed that strongly implies an automated system rather than a human recruiter made the decision [6]. The legal significance of the ruling is twofold. It signals that the company supplying the algorithm, not only the employer deploying it, may bear liability, which places the burden of demonstrating non-discrimination directly on the engineering vendor. It also shows how the pattern of the harm becomes the evidence: because the rejections were fast, numerous, and unexplained, the plaintiff could argue that a biased automated tool was responsible, and the vendor had no explanation trail to rebut the inference. Had the pipeline produced per-decision attribution records of the kind SHAP and LIME generate, the vendor could have shown which job-relevant features drove each rejection; the absence of that record is part of what allowed the claims to survive a motion to dismiss. The case is thus a concrete demonstration that unexplained automated decisions are becoming a direct legal liability.

Second, regulation. Regulation (EU) 2024/1689, the EU AI Act, entered into force in August 2024 and explicitly designates both employment-related and creditworthiness-assessment AI systems as high-risk under Annex III [5]. High-risk classification is not a label; it triggers a concrete set of binding obligations. Providers must establish a risk-management system, maintain detailed technical documentation, keep automatic logs of the system's operation, ensure a level of transparency that lets deployers interpret the output, and design the system so that a human can meaningfully oversee and, where necessary, override its decisions [5]. These obligations phase in over a multi-year horizon, with the high-risk provisions of Annex III taking effect through the 2026 to 2027 period, and non-compliance carries penalties that scale with global turnover [5]. The importance of this for the present report is that transparency, logging, and human oversight are precisely the artefacts that post-hoc explanation methods produce: an attribution profile is a form of documentation, an interpretable output, and a basis for human override all at once. Because European rules tend to become a global baseline through the market incentive to build a single compliant product, these requirements shape engineering practice well beyond Europe. Third, municipal law: New York City's Local Law 144 already requires annual independent bias audits of automated employment decision tools and public disclosure of the results, giving hiring-tool vendors an immediate, local reason to adopt the aggregate auditing that SHAP profiles support [11]. Together, these developments mean that the inability to explain an algorithmic decision is no longer merely a design weakness; it is a legal and financial liability.

5.3 Scope, Discipline, and Structure of This Report

This report is written from the perspective of computer engineering and applied artificial intelligence, and targets a general university-level engineering audience. It examines one challenge (algorithmic bias in automated hiring and credit-scoring decisions), one technology (post-hoc explainable AI, represented by SHAP and LIME), and one discipline (computer engineering and applied AI). Section 6 describes where bias enters automated decision pipelines and explains the mechanics of SHAP and LIME. Section 7

discusses how these methods mitigate bias in practice, evaluates their strengths and limitations, and situates them against regulatory requirements. Section 8 concludes with recommendations.

6. Main Topics

6.1 Where Bias Enters an Automated Decision Pipeline

Bias in machine-learning systems is rarely the result of malicious design; it is a property of the pipeline. Albaroudi, Mansouri, and Alameer's review of AI hiring systems identifies bias entry points at every stage: data collection (historical and sampling bias, as in the Amazon case), feature selection (proxy variables such as postal code or university name that correlate with protected attributes), model training (objective functions that optimize accuracy on majority groups), and deployment (feedback loops in which biased outputs generate further biased training data) [2]. Mehrabi et al.'s widely cited survey formalizes this taxonomy, distinguishing more than twenty mechanisms by which bias propagates from data to decisions [9]. Figure 1 summarizes these entry points and shows where the technology proposed in this report intervenes.

Figure 1. Bias entry points across an automated decision pipeline, and the post-hoc XAI audit layer (SHAP/LIME) that intervenes at the training-evaluation and deployment stages. Diagram by the author, synthesized from [2] and [9].

The critical observation from Figure 1 is that post-hoc explainability does not prevent bias from entering the pipeline; it makes bias visible at the output, where it can be detected, documented, and acted upon. This positioning shapes both the strengths and the limitations discussed in Section 7.

6.2 The Technology: Post-Hoc Explainable AI

Explainable AI refers to methods that make the behaviour of machine-learning models intelligible to humans. Two families exist: intrinsically interpretable models (such as small decision trees or linear scorecards), and post-hoc methods that explain an already trained black-box model. Because the accuracy advantages of gradient-boosted ensembles and deep networks are precisely what drives their adoption in hiring and lending, the practically relevant family is post-hoc explanation, and within it two techniques dominate real-world use: LIME and SHAP [4].

6.2.1 LIME: Local Interpretable Model-Agnostic Explanations

LIME, introduced by Ribeiro, Singh, and Guestrin in 2016, explains an individual prediction by probing the black-box model in the neighbourhood of the instance being explained. It generates perturbed copies of the input, queries the model on each, weights the results by proximity to the original instance, and fits a simple, interpretable surrogate model, typically a sparse linear model, to approximate the black box locally [8]. The

coefficients of the surrogate indicate which features pushed the specific decision toward acceptance or rejection.

More formally, LIME approximates the complex model in the region around a single data point by minimizing a weighted loss between the surrogate and the black box over a set of perturbed samples, subject to a complexity penalty that keeps the surrogate simple enough for a human to read. The proximity weighting is what makes the explanation local: samples close to the instance being explained count for more, so the surrogate captures the model's behaviour in that immediate vicinity rather than globally. The practical consequence is that a LIME explanation is a statement about one decision, not about the model as a whole, and two applicants with similar profiles may receive slightly different explanations. LIME is model-agnostic and computationally cheap, which explains its rapid adoption, but its explanations depend on the random perturbation sample, the kernel width that defines “nearby,” and the number of features retained in the surrogate, so repeated runs can yield different explanations for the same decision [4]. This instability is the principal reason auditors treat LIME as a fast diagnostic rather than as court-ready evidence.

6.2.2 SHAP: SHapley Additive exPlanations

SHAP, introduced by Lundberg and Lee in 2017, grounds feature attribution in cooperative game theory. It treats each input feature as a “player” in a game whose payout is the model's prediction and assigns each feature its Shapley value: its average marginal contribution across all possible orderings in which features could be added [7]. The intuition behind the Shapley value is a fairness argument borrowed from economics. To decide how much credit a feature deserves, SHAP imagines building the prediction up one feature at a time and measures how much that feature changes the output, then averages this marginal contribution over every possible order in which the features could have been introduced. Averaging over all orderings is what makes the attribution fair: no feature is arbitrarily privileged by being considered first or last.

This construction gives SHAP two properties that LIME lacks. The first is local accuracy, meaning the individual feature attributions sum exactly to the difference between the model's prediction for that instance and the baseline (average) prediction, so nothing is left unexplained. The second is consistency, meaning that if a model changes so that a feature contributes more to the outcome, that feature's attribution can never decrease. These guarantees are precisely why SHAP outputs are defensible under scrutiny: an auditor or a court can rely on the attributions summing correctly and behaving monotonically, rather than on the sampling luck that governs a LIME explanation. Salih et al.'s 2025 comparative perspective concludes that these guarantees, along with efficient implementations such as TreeSHAP for ensemble models, have made SHAP the most widely adopted XAI method in practice, using GitHub adoption metrics as evidence [4]. The main cost is computational: computing exact Shapley values requires evaluating the model over every subset of features, which grows exponentially with the number of features, so practical deployments rely on approximations such as KernelSHAP and TreeSHAP. These approximations introduce assumptions, most

notably feature independence, that real hiring and credit data frequently violate, and practitioners must understand those assumptions to avoid misreading the results [4]. The distinction between the two methods is therefore not merely academic: it determines which tool is appropriate for a fast, applicant-facing explanation (LIME) and which is appropriate for a rigorous, documented audit (SHAP).

Table 1 summarizes the comparison between the two methods.

Criterion	LIME [8]	SHAP [7]
Theoretical basis	Local surrogate model fitted to perturbed samples.	Shapley values from cooperative game theory.
Formal guarantees	None; explanation quality depends on sampling and neighbourhood choice.	Local accuracy, consistency, and missingness [7].
Stability	Can vary between runs on the same instance [4].	Deterministic for exact and tree variants; approximations add assumptions [4].
Computational cost	Low; fast per-instance explanations.	Higher; efficient for tree ensembles (TreeSHAP), costly for arbitrary models.
Typical audit role	Quick per-decision checks and applicant-facing explanations.	Regulatory audits, global bias profiling, documentary evidence.

Table 1. Comparison of LIME and SHAP as post-hoc explanation methods, synthesized from [4], [7], and [8].

6.3 What an Explanation Looks Like in a Lending Decision

To make the technology concrete, Figure 2 shows an illustrative SHAP-style attribution profile for a hypothetical declined credit application. Each bar quantifies how much a feature moved the model's output away from (dark) or toward (light) approval relative to the population baseline. Two aspects of such a profile matter for fairness work. First, it gives the applicant a specific, contestable explanation, such as "your credit-utilization ratio was the dominant factor," rather than a generic rejection, directly serving the transparency obligations the EU AI Act imposes on high-risk systems [5]. Second, and more important for auditing, it can surface proxy discrimination: in Figure 2, a nominally neutral feature such as ZIP code contributes negatively, a classic red flag because geographic features frequently encode protected attributes such as race [9]. Aggregating such profiles across thousands of decisions lets an auditor test statistically whether proxies drive systematically worse outcomes for protected groups, the same style of inference that supported the plaintiff's allegations in Mobley v. Workday [6].

Figure 2. Illustrative SHAP-style feature-attribution profile for a declined credit application (values are synthetic, for demonstration). The ZIP-code bar illustrates how attribution exposes potential proxy discrimination. Chart by the author, based on the attribution format of [7].

7. Discussion

7.1 How XAI Mitigates Bias: Three Mechanisms

The evidence reviewed above supports three distinct mitigation mechanisms. First, audit and detection: attribution profiles aggregated across decisions reveal whether protected attributes or their proxies materially influence outcomes, enabling the annual bias audits that New York City already mandates for hiring tools [11] and the conformity assessments the EU AI Act requires for high-risk systems [5]. Second, individualized transparency: per-decision explanations satisfy the growing legal expectation that affected individuals receive meaningful information about automated decisions, converting the applicant's experience from an unexplained rejection into a contestable one [3], [5]. Third, documentary accountability: attribution outputs create an evidence trail that vendors and deployers can produce in litigation or regulatory review; the absence of exactly this kind of explainability is what transformed the allegations in Mobley v. Workday from anecdote into a plausible legal claim that survived a motion to dismiss [6].

7.2 Strengths of the Proposed Approach

The principal strength of post-hoc XAI is that it decouples fairness auditing from model architecture. Because SHAP and LIME are model-agnostic, an institution can retain the accuracy of its gradient-boosted or deep model while adding an audit layer, rather than reverting to less accurate interpretable models, resolving the accuracy-transparency trade-off that Bahlool, Hewahi, and Elmedany identify as the central tension in credit-scoring research [3]. SHAP's game-theoretic guarantees give its outputs a principled, defensible basis, which matters when explanations must withstand regulatory or adversarial legal scrutiny [7]. The methods are also mature: both have open-source implementations with large practitioner communities, and Salih et al. document their dominance in applied deployments across domains [4]. Finally, the approach aligns with, rather than resists, the regulatory direction of travel: the transparency, logging, and human-oversight obligations of the EU AI Act are operationally achievable with exactly the artefacts SHAP and LIME produce [5].

7.3 Limitations and Risks

The limitations are equally clear from the literature, and an honest proposal must state them. Post-hoc methods explain; they do not repair. A perfectly explained biased model is still biased, so explainable AI must be paired with data-level and training-level interventions such as reweighting, debiasing, or fairness-constrained optimization [2], [9]. Explanation quality is also imperfect: LIME's outputs can be unstable across runs, and SHAP's practical approximations assume feature independence that real hiring and credit data violate, particularly under strong feature correlation [4]. There is, further, a human-factors risk: Salih et al. emphasize that end-users frequently misinterpret attribution values, for example reading them as causal effects rather than model-relative contributions, which can create unwarranted confidence in an audited-but-still-biased system [4]. Finally, there is an economic gap: the 2026 credit-scoring review notes that smaller

lenders without in-house data-science teams may struggle to implement joint performance-fairness-explainability pipelines, meaning regulation may consolidate advantage among large institutions [3]. Table 2 maps the emerging regulatory obligations to the specific XAI capabilities and residual gaps identified in this report.

Regulatory obligation	XAI capability that addresses it	Residual gap
Transparency to affected individuals (EU AI Act, Annex III high-risk) [5]	Per-decision attribution profiles (Fig. 2) stating dominant factors.	Lay users may misread attributions as causal claims [4].
Independent bias audits of hiring tools (NYC Local Law 144) [11]	Aggregated SHAP profiles across decisions to detect proxy influence.	Detection does not remove bias; requires data or model intervention [2], [9].
Human oversight and logging of high-risk decisions [5]	Explanations give reviewers a concrete basis to override outputs.	Oversight quality depends on reviewer training and incentives.
Litigation defensibility (e.g., Mobley v. Workday) [6]	Documentary evidence trail of decision factors.	Evidence can equally expose liability if bias is present.

Table 2. Mapping of emerging regulatory and legal obligations to XAI capabilities and residual gaps. Compiled by the author from [2], [4], [5], [6], [9], [11].

7.4 Synthesis: A Necessary Layer, Not a Complete Solution

Read together, the hiring-side and lending-side literatures converge on the same conclusion from independent directions. Albaroudi et al. conclude that purely technical fixes cannot resolve bias rooted in historical social data and call for human-centred XAI combined with regulation [2]; Bahlool et al. find that fairness and explainability are treated as compliance add-ons rather than co-designed properties and call for their deliberate integration [3]. This report's proposal operationalizes that convergence: deploy SHAP as the primary audit instrument (for its consistency guarantees and suitability for the tree-ensemble models common in tabular hiring and credit data), use LIME for fast per-decision, applicant-facing explanations, and embed both inside a governance loop of human oversight, standardized fairness metrics, and periodic revalidation. This addresses danah boyd's original framing of the challenge [1]: explainable AI does not by itself define what “fair” means, but it produces the shared, inspectable evidence that lets engineers, regulators, and affected communities negotiate that definition on common ground.

8. Conclusion

This report examined a grand challenge identified by expert danah boyd in BBC Future's survey of twenty-first-century challenges: high-stakes decisions are increasingly delegated to data-driven systems whose fairness is demanded but undefined and unverifiable [1]. Focusing on hiring and credit scoring, two

domains where the harms are documented in peer-reviewed syntheses [2], [3], demonstrated in practice [10], actively litigated [6], and now formally regulated as high-risk [5], [11], the report proposed post-hoc explainable AI, specifically SHAP and LIME, as a mitigation strategy within computer engineering and applied artificial intelligence.

The central argument is that explainability converts opaque predictions into auditable evidence. SHAP's game-theoretically grounded attributions support rigorous bias audits and regulatory documentation, while LIME provides fast, individualized explanations to affected applicants. The report is equally explicit about limits: post-hoc methods expose bias rather than remove it, their approximations can mislead under correlated features, and their outputs can be misinterpreted by non-expert users. Explainable AI is therefore proposed as a necessary audit layer within a broader socio-technical pipeline that also includes data-level debiasing, fairness-constrained training, trained human oversight, and compliance with emerging law.

For the engineering profession, the practical takeaway is that explainability has shifted from a desirable feature to a design requirement: under the EU AI Act's phased high-risk obligations and precedents such as Mobley v. Workday, the ability to answer the question “why did the system decide this?” is becoming as fundamental to automated decision systems as accuracy itself. Engineers entering this field should treat attribution tooling, fairness metrics, and audit documentation as core components of the system architecture, not as afterthoughts appended for compliance.

References

[1] B. Lufkin, “50 grand challenges for the 21st century,” BBC Future, Mar. 31, 2017. [Online]. Available: https://www.bbc.com/future/article/20170331-50-grand-challenges-for-the-21st-century. [Accessed: Jul. 8, 2026].

[2] E. Albaroudi, T. Mansouri, and A. Alameer, “A comprehensive review of AI techniques for addressing algorithmic bias in job hiring,” AI, vol. 5, no. 1, pp. 383 to 404, 2024, doi: 10.3390/ai5010019.

[3] R. Bahlool, N. Hewahi, and W. Elmedany, “Performance, fairness, and explainability in AI-based credit scoring: A systematic literature review,” Journal of Risk and Financial Management, vol. 19, no. 2, Art. no. 104, 2026, doi: 10.3390/jrfm19020104.

[4] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, and G. Menegaz, “A perspective on explainable artificial intelligence methods: SHAP and LIME,” Advanced Intelligent Systems, vol. 7, no. 1, Art. no. 2400304, 2025, doi: 10.1002/aisy.202400304.

[5] European Commission, “AI Act,” Shaping Europe's Digital Future, Aug. 2024. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai. [Accessed: Jul. 8, 2026].

[6] Mobley v. Workday, Inc., 740 F. Supp. 3d 796 (N.D. Cal. 2024).

[7] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Proc. 31st Int. Conf. Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 2017, pp. 4768 to 4777.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’: Explaining the predictions of any classifier,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 1135 to 1144, doi: 10.1145/2939672.2939778.

[9] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” ACM Computing Surveys, vol. 54, no. 6, Art. no. 115, 2021, doi: 10.1145/3457607.

[10] J. Dastin, “Amazon scraps secret AI recruiting tool that showed bias against women,” Reuters, Oct. 10, 2018. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. [Accessed: Jul. 8, 2026].

[11] New York City Department of Consumer and Worker Protection, “Automated Employment Decision Tools (Local Law 144 of 2021),” NYC.gov. [Online]. Available: https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page. [Accessed: Jul. 8, 2026].

Appendix A: Revision Summary

This appendix summarizes the feedback I received through the PeerScholar Assess phase and the changes I made to my Phase 2 draft in producing this final report. I was assigned three peer reviewers. Because reviewing is blind, two of my reviewers (Peers 1 and 2) were evaluating reports on other topics (a concentrated-solar-power proposal and an autonomous-vehicle sensor-fusion proposal), so parts of their commentary described those documents rather than mine. Where their advice was general and transferable, however, I applied it to my own report, and I note below both what I changed and why some comments did not apply.

The most directly relevant feedback came from Peer 3, who assessed my report at the level the assignment asks for: proofreading, sentence structure, and paragraph structure and flow. Peer 3's central points were that I should confirm every claim is cited, verify that no section was lost on export, and adopt a more consistently formal tone. I agreed with these points and treated them as the core of my revision. The table below maps each item of feedback to the specific change I made.

Feedback received (PeerScholar)	Change made in the final report
Peer 1: Add in-text citation numbers next to specific claims so readers can trace them.	Verified that every substantive claim carries a bracketed IEEE citation and added citations to a small number of sentences in the Discussion that previously relied on an earlier reference.
Peer 1 and Peer 3: Add a system diagram and a comparison table to clarify the technical sections.	Kept and refined Figure 1 (pipeline) and Table 1 (LIME vs SHAP), and confirmed each is introduced and interpreted in the surrounding text rather than left to stand alone.
Peer 2: Reference [1] linked to a forum mirror rather than the official BBC page.	Confirmed reference [1] points to the official BBC Future URL and re-checked every hyperlink and access date in the reference list.
Peer 2: Add a rough regulatory or deployment timeline for staged adoption.	Expanded the regulatory discussion with the EU AI Act phase-in horizon (in force 2024; Annex III high-risk obligations phasing through 2026 to 2027) and retained the obligation-mapping Table 2.
Peer 3: Ensure required report sections are all present and that no section is truncated on export.	Re-exported the document and verified all ten technical-report sections are present and complete, with Letter of Transmittal, Title Page, Executive Summary, Conclusion, and References each on their own page.
Peer 3: Adopt a more formal tone with fewer contractions.	Removed remaining contractions and informal phrasing and tightened wording for a consistent formal register aimed at a university engineering audience.

Counter-arguing and Reconciling Feedback

Not all feedback pointed the same way, and one suggestion I respectfully set aside. Peer 2 recommended adding a dollar cost estimate for staged deployment. Because my proposal concerns a software auditing method (SHAP and LIME are open-source libraries that attach to an existing model) rather than a capital-intensive physical system, a monetary cost estimate would be speculative and potentially misleading. Instead, I addressed the underlying concern, which is feasibility and timeline, by strengthening the regulatory timeline in Section 5.2 and by retaining Table 2, which maps concrete obligations to the capabilities that satisfy them. This honours the intent of the comment while keeping the report accurate.

The feedback did not otherwise contradict itself: Peers 1 and 3 both pushed toward clearer citations and stronger visuals, which reinforced rather than opposed each other, so no genuinely contradictory advice needed reconciling. I also went beyond the peer comments, as the assignment encourages, by verifying pagination so that the Letter of Transmittal, Title Page, Executive Summary, Conclusion, and References each occupy their own page, and by confirming that the main body (Sections 1 through 4) meets the required length.

Reflection: Final Report versus Draft, Including Unedited Sections

Compared with the Phase 2 draft, the final report is tighter in tone, fully cited at every claim, and clearer about how its two figures and two tables support the argument. The substantive technical content of Sections 2 and 3 was already strong in the draft and was deliberately left largely unedited, because the peer feedback did not identify weaknesses in the mechanics of SHAP and LIME or in the evidence base, and changing accurate, well-supported material for its own sake would have risked introducing errors. The reference list was likewise retained in full, since it already drew on eleven sources including four peer-reviewed articles and the two foundational SHAP and LIME papers, exceeding the requirement of at least two peer-reviewed sources. The most meaningful improvements were therefore at the level of citation completeness, tone, visual integration, and document formatting, which is precisely where the peer feedback was directed.

Appendix B: AI Use Disclosure Statement

As far as my work on this project goes, my use of generative AI was mainly limited to its use for research purposes and improving the quality of the content. I used AI to verify grammar, correct minor mistakes, and to enable me to convey some complicated concepts related to the topic, like the Shapley-value foundation of SHAP versus the local surrogate foundation of LIME in simple terms to an audience of engineers. Additionally, I have used AI to produce the sample attribution table seen in Figure 2 to illustrate the point in a clearer manner, specifying the features and their synthetic values by myself. The reason why I could evaluate the recommendations of the AI critically is because I had previous experience in the subject matter.

As for my contribution, I have done much more than just using the AI service. I chose all the sources myself, checked every citation and access date against the original, formed an argument, decided what claims are supported by each of the sources and made the engineering conclusions in the analysis. In addition, after producing drafts of the passages, I edited them considerably. I fixed all technical claims, got rid of the unsupported claims, made the tone more formal according to my peers' suggestions, and produced the revision summary and this declaration myself.