ENG 2003 - Effective Engineering Communication

Term Project 01: Phase 1 – Annotated Bibliography

Name: Ariyan Azami

Student ID: 222004071

Topic Statement

Challenge:	**danah boyd, Principal Researcher at Microsoft Research, names as one of the grand challenges of the 21st century the fact that data-driven technologies are increasingly embedded in high-stakes decisions, from judicial rulings to hiring and lending, and that regulators, civil society, and social theorists increasingly demand these systems be "fair" and "ethical," even though those concepts remain poorly defined and hard to translate between technical and social actors [B. Lufkin, "50 grand challenges for the 21st century," BBC Future, Mar. 31, 2017]. This connects directly to my topic: as automated decision systems mediate hiring and lending outcomes, and as deep learning has improved predictive performance while reducing transparency, explainable AI methods such as SHAP and LIME are proposed as one way to make these systems' fairness and accountability legible to regulators, auditors, and the people affected by them.**
Technology (Solution):	Explainable AI (XAI), including post-hoc methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), used as a bias-mitigation layer in automated decision systems.
Area of Interest (Scope or Discipline):	Computer engineering and applied artificial intelligence, carried forward from Assignment 3 (algorithmic bias in hiring/lending and XAI as a mitigation strategy).
Topic Title OR Statement:	Algorithmic Fairness in Automated Decision Systems: Explainable AI as a Mitigation Strategy for Bias in Hiring and Credit-Scoring Pipelines

*Annotated Bibliography Section – refer to the Term Project 1 Handout **&** Phase 1 Grading Rubric for exact details of what's required in Phase 1.*

Table 1: Annotated bibliography writing steps for each research source

Summarize (20% of grade)****	**Provide the source's citation using the IEEE Style (format). Briefly summarize the content of each reference (150-200 words). What is the purpose of the work? What was the methodology followed? Try to define the intended audience of the work.**
Evaluate** (30% of grade)**	What are the strengths of this cited work? Think back to rhetorical triangle (ethos/credibility, logos/logic and evidence, and pathos/audience-centric) and deliberative rhetoric (future persuasion). What are the weaknesses of the work in each reference? Provide a critical review of the presented idea, used method, and outcomes.
Reflect (50% of grade)	What interested you in this source? How did you come across this source? How is it related to your topic of term project 1? What impact will this work have on your topic and field of study? Do you think this source will still be relevant another 5 years from now?

Source 1

Title & Year	A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring (2024)
Complete IEEE Citation	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. 383–404, 2024, doi: 10.3390/ai5010019.
Summarize	This peer-reviewed review, published in the MDPI journal AI, synthesizes research on algorithmic bias in AI-driven hiring systems, covering resume screening, automated video interviews, and personality-assessment tools such as HireVue and Pymetrics. Its purpose is to catalogue where bias enters a hiring algorithm's pipeline (data collection, feature selection, training, deployment) and to review technical mitigation strategies, including explainable AI, proposed to counter it. Methodologically, the authors conduct a structured literature review, aggregating findings across computer science, HR, and legal scholarship, a cross-disciplinary lens that is unusual for a technical AI paper and strengthens its relevance to my project. The intended audience is dual: AI researchers designing fairness-aware hiring tools, and HR or compliance professionals who need to understand the technical vocabulary without a computer-science background. The paper concludes that human-centered XAI design and stronger regulatory frameworks are both necessary, since purely technical fixes cannot resolve bias that is rooted in historical, social data.
Evaluate	Strengths: high ethos, since the paper is peer-reviewed and published in an indexed, open-access AI journal, and its systematic-review methodology (aggregating dozens of primary studies) gives it strong logos, its claims rest on triangulated evidence rather than a single case. It also engages pathos by centering the applicant's lived experience, echoing earlier HCI critiques of algorithmic scoring as "reducing a human being to a percentage." Weaknesses: as a review rather than original empirical work, it inherits the limitations of the studies it summarizes and contributes few new quantitative results of its own; its recommendations (human-centered XAI, stronger regulation) are somewhat generic and under-specify how engineers should implement them in practice. In deliberative-rhetoric terms, the paper is forward-looking about what hiring AI should become, but stops short of a concrete design roadmap, which limits its persuasive force as a call to action.
Reflect	I found this source while searching for recent scholarly work matching the exact phrase "algorithmic bias in job hiring," since it directly aligns with my group's Assignment 3 topic. It anchors the "Challenge" side of my project by mapping where bias enters a hiring pipeline, letting me argue later, in the Phase 2 report, that XAI methods such as SHAP and LIME intervene specifically at the post-processing/audit stage rather than fixing the pipeline end to end. Because it is a 2024 synthesis of dozens of studies rather than a single vendor's case study, I expect it to remain a useful reference point five years from now, even if individual tools it discusses (HireVue, Pymetrics) are eventually superseded, since systematic reviews document a problem space rather than one product's behavior.

Source 2

Title & Year	Performance, Fairness, and Explainability in AI-Based Credit Scoring: A Systematic Literature Review (2026)
Complete IEEE Citation	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.
Summarize	Published in the Journal of Risk and Financial Management, this systematic literature review examines 43 peer-reviewed studies from 2020–2025 on AI-based credit scoring, addressing predictive performance, fairness, and explainability together. Its purpose is to assess whether existing credit-scoring research adequately balances accuracy with fairness and interpretability under regulatory oversight, rather than treating these three dimensions as separate research problems. Methodologically, the authors follow a formal systematic-review protocol, searching IEEE Xplore, Scopus, Web of Science, and ScienceDirect for studies published between 2020 and 2025, then grading each included study with a "3Rs&Q" (Relevance, Rigor, Reproducibility, and Quality) appraisal framework, a more rigorous and transparent methodology than an ordinary narrative review. The intended audience is academic and industry researchers in financial technology, as well as regulators evaluating AI-governance frameworks for lending decisions. Its central finding, that performance, fairness, and explainability are usually addressed in isolation rather than jointly, directly supports my project's argument that XAI needs to be deliberately integrated as a bias-mitigation layer rather than added afterward as a compliance patch.
Evaluate	Strengths: very strong logos, the formal systematic-review protocol (structured screening plus the 3Rs&Q framework) gives the paper high methodological credibility, and because it was published in early 2026 it captures the current state of the field, including SHAP-based feature-attribution frameworks for credit models. Its ethos also benefits from appearing in a specialized, peer-reviewed finance journal rather than a general AI venue. Weaknesses: as a review of others' work, it cannot independently verify whether the fairness metrics reported by the underlying 43 studies were computed consistently, so its conclusions rely on the quality of results reported elsewhere. It is also largely silent on how smaller lenders without in-house data-science teams could realistically adopt the joint performance-fairness-explainability approach it recommends, a practicality gap common in academic surveys. Its deliberative framing pushes for future integration research but does not yet demonstrate feasibility at scale.
Reflect	I looked specifically for recent (2025–2026) research on the credit-scoring side of my topic, since our group's outline covers both hiring and lending decisions. This source is directly relevant because it independently confirms, from the finance side, the same gap the hiring source above identifies from the HR side: fairness and explainability tend to be treated as separate add-ons rather than co-designed with model performance. This strengthens the central claim of my Term Project 1 report, that a single mitigation strategy, XAI, can be discussed meaningfully across both domains. Given that it was published in February 2026 and that the EU AI Act's high-risk provisions for credit scoring are still being phased in through 2026–2027, I expect this source to remain relevant for several years as regulatory deadlines approach and institutions scramble to comply.

Source 3

Title & Year	A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME (2024/2025)
Complete IEEE Citation	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.
Summarize	This peer-reviewed perspective paper, published in Wiley's Advanced Intelligent Systems, directly compares the two most widely used post-hoc explainability techniques: SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). Its purpose is to give practitioners a clear framework for choosing between the two methods and interpreting their output correctly, since the authors note that end-users often misunderstand the assumptions behind each technique. Methodologically, the authors compare SHAP and LIME both theoretically, SHAP is grounded in cooperative game theory and Shapley values, while LIME fits a local surrogate model, and empirically through case studies, using GitHub star counts as a proxy for real-world adoption. The intended audience is applied machine-learning practitioners and domain experts (the case studies are drawn from biomedicine) who need to select and correctly interpret an XAI method, rather than AI theorists. This paper is the technical backbone of my project, since SHAP and LIME are the two specific technologies named in our group's Assignment 3 outline as the bias-mitigation tools for hiring and credit-scoring pipelines.
Evaluate	Strengths: strong logos, since the comparison is both theoretically grounded (game theory vs. local surrogate modeling) and empirically demonstrated through real case studies, and the GitHub-star adoption metric is a clever, practitioner-relevant proxy for real-world credibility. Ethos is high given the multi-institutional, peer-reviewed authorship and the journal's Scopus/Web of Science indexing. Weaknesses: the case studies are drawn from biomedical data rather than hiring or credit-scoring contexts, so I have to extrapolate its conclusions about model-dependency and feature collinearity to my own domain rather than relying on a direct precedent. It is also a "perspective" piece rather than a full empirical benchmark study, so its comparative claims, while well-reasoned, are less exhaustively validated than a large-scale study would be. In terms of deliberative rhetoric, the paper persuasively frames explainability as something end-users must learn to interpret correctly, not just something engineers must produce, a genuinely forward-looking argument about responsible XAI adoption.
Reflect	I found this paper while researching the exact SHAP/LIME comparison referenced in our tutorial outline, since our group needed a citable, technical source explaining what these two methods actually do and how they differ. It is directly related to my Term Project 1 topic because SHAP and LIME are the specific "Technology" solution our group proposed for mitigating hiring and lending bias, so this source lets me explain the mechanics of the technology accurately rather than relying on informal descriptions. I believe this source will still be highly relevant in five years, since SHAP and LIME remain, by the paper's own evidence, the two dominant XAI methods in practice, and Shapley-value-based explanation is increasingly being written directly into regulatory guidance, including under the EU AI Act's transparency articles discussed below.

Source 4

Title & Year	AI Act | Shaping Europe's Digital Future (European Commission, in force since Aug. 2024)
Complete IEEE Citation	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. 5, 2026].
Summarize	This is the European Commission's official policy page describing Regulation (EU) 2024/1689, the AI Act, the world's first comprehensive horizontal legal framework for artificial intelligence. Its purpose is to inform providers, deployers, and the public about the Act's risk-based structure and their compliance obligations. Unlike the scholarly sources above, this is a primary legal and government source rather than an empirical study; its "methodology" is legislative documentation and official guidance rather than research design. Both hiring and employment tools and credit-scoring systems are explicitly designated "high-risk" under Annex III of the regulation, subjecting them to mandatory transparency, human-oversight, and documentation requirements. The intended audience is broad: AI developers and deployers who must achieve compliance, regulators enforcing the Act, and members of the public whose employment or credit opportunities are affected by these systems. For my project, this source supplies the regulatory "why it matters" context that motivates the technical discussion of explainable AI as a mitigation strategy.
Evaluate	Strengths: maximum ethos, since it is the official primary source from the regulating body itself rather than secondary commentary; none of my other sources carries as much authoritative weight regarding what companies are legally required to do. Its logos is strong in the narrow sense that legal text is precise and enforceable, with defined penalty structures reaching up to seven percent of global turnover. Weaknesses: the page is descriptive rather than analytical; it does not weigh the costs and benefits of the regulation or engage with counterarguments, such as industry concerns about compliance burden delaying deployment, so it offers ethos and information but little of the argumentative logos or pathos found in a scholarly or journalistic source. It is also a living document: implementation deadlines have already shifted once, with Annex III high-risk obligations proposed to move from August 2026 toward December 2027, so any claim drawn from it needs an access date, which I have included in the citation.
Reflect	I sought out this source specifically because our group's outline references the EU AI Act as a driver of why explainability is shifting from a desirable feature to a regulatory necessity, and I wanted that claim grounded in the Act's own text rather than a secondhand paraphrase. It connects directly to my project's argument that the pressure to adopt XAI in hiring and credit scoring is not just an engineering best practice, it is becoming a binding legal requirement in a major global market. Given that the AI Act's high-risk obligations are still being phased in through 2026 and beyond, and that EU rules tend to become a de facto global standard through the so-called "Brussels effect," I expect the regulation this page describes to remain highly relevant well beyond five years, even as the page's specific content is updated over time.

Source 5

Title & Year	Mobley v. Workday, Inc. (2024)
Complete IEEE Citation	Mobley v. Workday, Inc., 740 F. Supp. 3d 796 (N.D. Cal. 2024).
Summarize	This is the U.S. District Court for the Northern District of California's July 2024 ruling in Mobley v. Workday, one of the first major court decisions to address whether a third-party AI hiring-software vendor can be held directly liable for employment discrimination. The purpose of the opinion is to resolve a motion to dismiss: whether Workday's algorithmic applicant-screening tools made it an "agent" of the employers that used them under Title VII, the ADEA, and the ADA. The "methodology" is legal reasoning applied to the plaintiff's factual allegations, including that Derek Mobley, an African American man over 40 with a disability, was rejected from over 100 positions, sometimes within an hour of applying, an inference the court found sufficient to let the discrimination claims proceed to discovery. The intended audience is legal practitioners, HR-technology vendors, and employers, though its facts have since been widely reported to a lay audience as a landmark AI-bias case. It gives my project real-world case evidence to substantiate the abstract algorithmic-bias claims made in the scholarly sources above.
Evaluate	Strengths: extremely strong ethos and logos as a primary legal source, courts must justify every step of their reasoning against applicable statutes and precedent, and this opinion documents concrete, verifiable facts (application volume, near-instant automated rejections, alleged reliance on biased training data) that are more persuasive to a skeptical reader than an abstract description of "algorithmic bias." It provides exactly the kind of case-evidence our group's outline calls for, grounding the report's theoretical discussion in a real, ongoing dispute. Weaknesses: the opinion only resolves a motion to dismiss, meaning the court accepted the plaintiff's allegations as true for that motion; it is not a final finding that Workday's tools actually discriminated, so I must avoid overstating what has been legally proven. It is also fast-moving litigation, a further class-certification ruling followed in 2025, so any claim I make needs to note this is an evolving, unresolved case rather than settled law.
Reflect	I came across this case while researching real-world examples of AI hiring-bias litigation to complement the Amazon resume-screening story referenced in our tutorial outline, and found it more valuable because it is more recent and still actively developing, giving my report a live example rather than a purely historical one. It relates directly to my topic because it demonstrates exactly the "Case Evidence and Discussion" section our group planned, showing concretely how the absence of explainability in an AI hiring pipeline becomes a legal liability, precisely the gap XAI methods like SHAP and LIME are proposed to close. Because litigation of this kind is still unfolding, with class certification granted in 2025 and further proceedings likely, I expect this specific case to be even more relevant in five years than it is today, either as a landmark precedent or as a cautionary tale, depending on its outcome.