Le Manh Ha, 2025, 13:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

A Hybrid Machine Learning and XAI Architecture for Intelligent Career Guidance Systems

Le Manh Ha[1], Nguyen Huu Quynh[2], Nguyen Tai Tuyen[3]
[1]Faculty of Information Technology, Ha Long University, Quang Ninh, Vietnam
[2]CMC University, Hanoi, Vietnam
[3]Faculty of Electronics Engineering 1, Posts and Telecommunications Institute of Technology, Vietnam

Abstract- Industry 4.0 and artificial intelligence are shown to bring great change or disappear a large proportion of work, while new jobs are born With the dynamical background comes the demand for a smart career guidance system, which will provide advice that is personalized, reliable, and adaptive. This paper systematically reviews the literature on the application of machine learning (ML) and explainable artificial intelligence (XAI) to career guidance. On the basis of PRISMA guidelines, 847 documents published between 2019 and 2025 were carefully screened, and 95 high-quality articles were extracted for in-depth review. The review classifies main ML methods—including collaborative filtering, content-based filtering, deep learning architectures such as LSTM, Transformer, and GNN, reinforcement learning, and their performance, limitations, and interpretability were judged. In parallel, the author analyzes such essential but questioned XAI techniques as LIME, SHAP, attention mechanisms, decision rules and counterfactual explanations in terms of their transparency and perceived user trust, as well as how easily acted upon these explanations are. From these foundations, the paper presents a five-layer hybrid ML-XAI framework that integrates data processing, knowledge maps, ensemble ML models, multi-level explanations, and user-centered presentation. In addition to these, future developments, such as flat or formidable language models and federated learning for maintaining privacy and fairness-aware algorithms, are explored, together with key challenges for further research. All in all, the paper provides a structured basis and practical guidance for next-generation, intelligent, transparent, and equitable career guidance systems.

Keywords: Career Guidance, Machine learning, Explainable AI, Job recommendation, Hybrid ML-XAI Systems.

I. INTRODUCTION

The development of industry 4.0 and artificial intelligence has greatly affected the employment market. The urgent thing for people now is how to adapt to these changes and make decisions about what type of work they should have - or even if that question still exists! Based on this new technology, World Economic Forum (2023), predicts that fully 23% of today's jobs will experience significant changes in their duties within the next five years. Approximately 69 million new positions have yet to be generated and 83 million displaced. Workers need intelligent career guidance systems suited to this fluid world, capable of sifting through multi-dimensional data to help tailor advice that suits each person's skills and tastes while reflecting changing trends in the labor market at any given time.

In this environment Machine Learning (ML) has shown remarkable success. By analyzing great amounts of job data, user profiles and trends for people's paths into new careers it can provide personalized suggestions based on performance indicators gleaned from all sources. Studies also show that ML systems for decision making have achieved accuracies of 75-85% [1]. However, because advanced machine learning models, such as Deep Neural Networks, are essentially "black boxes," providing trustworthy and acceptable recommendations-especially those related with consequent effect–presents a challenge to developers.Explainable Artificial Intelligence (XAI) is proposed as a potential way to address the opacity law in AI systems.

Tools like LIME (Local Interpretable Model-agnostic Explanations) [2], SHAP (SHapley Additive exPlanations) [3] and attention mechanisms mean users can see why the system suggested what it did. Research shows that if explanations increased user confidence even just 20% they would be more likely to accept recommendations. In turn we received favorable response rates when introducing these to

our algorithms of 40%-60% on average.In this paper, we aim to achieve the following: (1) Analyze and categorize methods for ML-XAI career guidance by defining models able to work across architectures. (2) Assess hybrid architectures' overall effectiveness; (3) Spot challenges and research prospects; (4) Propose an integrated plan for the next generation as well as some basic principles of intelligent career guidance systems.

II. RESEARCH METHODOLOGY

A. Search Strategy

We followed the PRISMA guidelines for this study. Searching GB/T 4259:1984, IEEE Xplore, ACM Digital Library, ScienceDirect and Springer Link, Web of Science all produced results for searches. We also used Scopus and Nature Databases; during the part where we made search requests, which had been sent which of its elements would contribute most towards meeting that goal Search terms were combined with Boolean operators AND and OR.

B. Inclusion and Exclusion Criteria

Inclusion criteria: Recognized research journal during 2019 and 2025 year The article should engage in applications of AI and ML Career Counseling or Job Search Offer explanations of how trust can be verified Exclusion criteria: No full papers only conference abstracts Focus studies exclusively on campus recruiting is an enterprise-level perspective In Times with Mediocrity Papers that are merely literature reviews.

C. Screening Process

The screening process had three stages, namely title and abstract screening, full-text review and quality assessment. From the original 847 documents, 523 remained after duplicates were discarded. Meanwhile, the review of 189 papers following title and abstract screening led to only 95 passing full-text reviews that met all criteria for inclusion in this analysis. The distribution of papers coded according to: ML methods, XAI technology, data types, evaluation methods, and key results is shown in Worksheet 1.

III. MACHINE LEARNING APPROACHES IN CAREER GUIDANCE

A. Collaborative and Content-based Filtering

Collaborative Filtering (CF) is based on the premise that users' preferences may be very similar at times and so too will their career direction. Matrix factorization (MF) Multiple technology components are able to break down matrices of user-occupation interactions representing hidden attributes MF techniques into latent factors SVD++ extends SVD by incorporating browsing behavior and interaction with latent content MF-based CF of NDCG@10 ranging from 0.72 to 0.783 has been confirmed on large-scale job datasets by tests conducted by Qin et al. [5]'

Content-based Filtering (CBF) starts by comparing job descriptions and applicant profiles for a match. TF-IDF and Word2Vec are used to represent skills and requirements texts. Advanced semantic models such as BERT have improved contextual understanding of the contents of job descriptions. Hybrid systems combining CF and CBF services can deal with the cold-start problem, and they obtain an accuracy rate 10% to 15% higher than that of a single-method approach [6].

B. Deep Learning Architectures

Recurrent Neural Networks (RNN) user long and short term memory (LSTM) networks to model career history sequences and infer likely future directions LSTM EB uses the internal state of a given LSTM to learn a pattern from what has gone on before, and produce forecasts for what will be the next location. Bi-directional LSTM with lasso achieved Context recognition in the past and future career sequences of 82% Sun et al. [7] got a 3-LSTM model to offer career transition prediction accuracy rate of 0.2 miles between two points using the model and the historical point-data thus established as predictors.

Transformer architecture with built-in attention has overturned natural language processing in the career literature. BERT Bidirectional Encoder Representations from Transformers has been retrained specifically for job-skill matching and resume segmentation tasks. JobBERT And an

iteration that was prescaled on 10-billion-word corpora Ases a score of 0.9 for the skill extraction F1 measure Sentence-BERT Between job descriptions and candidate files, it provides efficient semantic comparison BERT Cross-Encoder It solves sentences in batch mode with 100x faster than any BERT model that does not deal with the parallel problem.

Graph Neural Networks (GNNs) leverage the relational structure of skill-occupation networks to capture richer semantics. GraphSAGE summarizes information from adjacent nodes, producing embeddings for users and jobs alike. JobEdKG (job education knowledge graph) models the relationships between skills, fields of study and occupations that are uncertain or even non-conventional, with just over 500 000 entity entries and 2 million relationships [9]. Graph Attention Networks (GAT) Attended Weights Learn for Edges, Letting the Model Focus on Important Relationships.

C. Reinforcement Learning for Personalized Paths

Reinforcement Learning (RL) is aimed at optimizing learning paths and skill acquisition. The RL model uses users as agents, the current state of skills as states, courses and careers as actions and rewards for career progress. Deep Q-Network (DQN) has learned to recommend courses in order to achieve career goals. Policy gradient methods like A2C can provide a direct policy instead of an appropriate learning something which is suitable for large-scale operational systems in the political domain. Zhou et al. [10] have shown that RL-based recommendations Improve Completion rate by 25% When Compared to Rule-Based Systems.

TABLE I
COMPARISON OF MACHINE LEARNING METHODS

Method	Advantages	Limitations	Accuracy	Interpretability
Collaborative Filtering	Discovers latent preferences; No domain knowledge required	Cold-start problem; Data sparsity	NDCG: 0.72-0.78	Low
Content-based	Feature-based explanations; No item cold-start	Over-specialization; Feature engineering dependent	Precision: 0.68-0.75	Medium
LSTM/RNN	Sequence modeling; Career trajectory	Vanishing gradient; Slow training	Accuracy: 78-82%	Low
BERT/Transformer	Semantic understanding; Pre-trained knowledge	High computational cost; Requires fine-tuning	F1: 0.85-0.92	Medium
GNN/Knowledge Graph	Rich relationships; Explainable paths	Complex graph construction; Scalability issues	MRR: 0.65-0.72	High
Reinforcement Learning	Long-term optimization; Dynamic adaptation	Reward design difficulty; Sample inefficiency	Reward +25%	Low

IV. EXPLAINABLE AI TECHNIQUES

A. Post-hoc Explanation Methods

LIME generates local explanations by approximating complex models with simple models (typically linear) in the neighborhood of the data point being explained. In career guidance contexts, LIME can identify which skills or experiences contribute most to a specific career recommendation. For example: "Data Scientist is recommended because: Python skill (+0.35), Machine Learning experience (+0.28), Master's degree (+0.15)." Zhang et al. [11] found that LIME explanations increase user trust by 35% in job matching systems.

SHAP is based on game-theoretic Shapley values to fairly allocate the contribution of each feature to the prediction. TreeSHAP is optimized for tree-based models such as XGBoost and Random Forest, commonly used in job recommendation. DeepSHAP combines SHAP with DeepLIFT for deep neural networks. SHAP provides both global importance (overall important features) and local explanations (explanations for individual recommendations). The advantage of SHAP is its mathematical consistency and ability to compare feature contributions across predictions [3].

B. Intrinsic Interpretability

Attention weights in Transformer models provide insights into which parts of the input the model

focuses on. Multi-head attention allows the model to learn multiple attention patterns simultaneously. Visualization of attention maps helps understand what the model is "looking at" when making decisions. In job-resume matching, attention can highlight which skills in a CV match job description requirements. However, attention weights do not always correspond to actual feature importance [12]. Decision Trees and rule extraction methods provide explanations in easily understandable "IF-THEN" format. Techniques such as RIPPER and C4.5 generate rule sets from training data. Rule extraction from neural networks uses techniques such as decompositional approaches. Example rule: "IF programming_skill >= 3 AND math_skill >= 4 AND experience_years >= 2 THEN recommend: Data Analyst." Hybrid models combine neural network accuracy with rule interpretability [13].

C. Counterfactual Explanations

Counterfactual explanations answer the question "What needs to change to get a different outcome?" In career guidance, counterfactuals indicate: "If you had 2 additional years of Python experience, you would be recommended for Senior Developer instead of Junior Developer." DiCE (Diverse Counterfactual Explanations) generates multiple diverse counterfactuals giving users multiple development options [14]. Counterfactuals are particularly useful for actionable recommendations—helping users know specifically what to do to achieve career goals.

TABLE II
COMPARISON OF XAI TECHNIQUES

Technique	Type	Scope	Advantages	Limitations
LIME	Post-hoc, Model-agnostic	Local	Flexible, intuitive output	Instability, sampling dependency
SHAP	Post-hoc, Model-agnostic	Local + Global	Theoretically grounded, consistent	High computational cost
Attention	Intrinsic	Local	Built-in, no extra computation	May not reflect true importance
Decision Trees	Intrinsic	Global	Highly interpretable rules	Limited accuracy for complex tasks
Counterfactuals	Post-hoc	Local	Actionable insights for users	May suggest infeasible changes
KG Paths	Intrinsic	Local + Global	Semantic relationship explanations	Graph construction overhead

V. HYBRID ML-XAI ARCHITECTURES

A. Integration Models

The complex ML model offers the prediction power, and the XAI brings a sense of reason. In which there are three main fusion models: (1) In-Line Integration-predictions from an ML model first are explained, and then explanations offered; (2) Integrated.gif XAI is part of the training process for the model after all; (3) Mixed or eclectic--Ensemble integration combines interpretable lower level models with single complex ones. Each method has advantages and disadvantages in terms of accuracy, interpretability and computational cost [15].

B. Proposed Framework

Taking a cue from these considerations, this study proposes a five-layer Hybrid ML-XAI framework: (1) The Data Layer collects and pre-processes data from various sources (user profiles, job descriptions, market trends); (2) The Knowledge Layer constructs a Knowledge Graph connecting skills, occupations, and industries. (3) The ML Layer is an ensemble of Deep Learning (BERT, GNN) plus traditional ML (CF, CBF). (4) The XAI Layer includes LIME, SHAP, attention visualization, counterfactuals. (5) The Presentation Layer is an interface showing recommendations with explanations customized for different user types.

This framework meets requirements for personalization with multi-modal User Profiling, accuracy through ensemble ML methods, transparency from multi-level explanations, and adaptability by continual learning from user feedback. Implementing guidelines include modular architecture for component updates, API-based integration to keep flexibility open when using Modules, and user study protocols to evaluate explanations.

C. Case Studies

LinkedIn Skills Graph: LinkedIn has built a knowledge graph of over 50,000 skills and their interrelationships. The system employs a GNN to embed skills and jobs, combined with collaborative filtering based on the career transitions for over 700 million users. Explanations are provided in "People with similar backgrounds also moved to..." as well as in The mind of its readerserves us up skill-gap analyses [16].

ESCO (European Skills/Competences, Qualifications and Occupations): ESCO provides a multilingual classification of skills, competences, qualifications, and occupations. Hybrid systems using ESCO ontology combined with ML models have been deployed at employment service agencies throughout the EU. Skill-job mappings that users understand are interpretable, are provided [17].

VI. LARGE LANGUAGE MODELS AND EMERGING TRENDS

A. LLM Applications in Career Guidance

Claude, Gemini and other Large Language Model applications now make career guidance systems interactive and allow natural generation of content. In this area, there are a few key uses: (1) Conversational career counseling--Integrated LLMs can ask questions and do interview tests to evaluate what kind of work people want or are fit for; (2) Resume production and fine-tuning-- they can produce CVs to match certain jobs and improve an existing one; (3) Job description generation-- they can make job descriptions that match the other side's personnel needs and position requirements; (4) Skill gap analysis--can identify who lacks which skills and what to study next. [18]

Li et al. [19] proposed the GIRL (Generative Job Recommendations with Large Language Model) framework, fine-tuning large language models using Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) for job recommendation. In evaluation, LLM-based approaches win 65-70% of matches against traditional methods when assessed by the relevance, detail level, and conciseness of generated job descriptions according to ChatGPT.

B. Federated Learning for Privacy Preservation

Federated Learning (FL) allows you to train models on distributed data without centralizing sensitive user information. FedRec frameworks for recommendation systems keep user interaction data locally on devices and only share model updates. FedRKG (Federated Recommendation with Knowledge Graph) uses the global KG on the server, and local user embeddings, to reach near-central levels of accuracy while still maintaining privacy [20]. Local Differential Privacy (LDP) is added so gradients sent cannot be reverse-engineered to reconstruct user data. Studies have shown that FL-based recommenders perform at 95-98% accuracy compared with their centralized counterparts.

VII. ALGORITHMIC FAIRNESS AND BIAS

A. Types of Bias in Job Recommendation

Algorithmic bias in recruitment and career guidance systems has been documented through multiple studies. Gender There have been a number of reports documenting algorithmic bias in recruitment and career guidance systems. Amazon discontinued its AI hiring tool due to gender bias in 2018. There, the system downranked any CVs containing "women's," a behavior traceable to the training data. For racial bias: University of Washington's 2024 LLM research finds that 85% of white-associated names are favored 100% in all experiments, while black male associated-name never enjoys favoritism [22] EEOC settled its first AI age discrimination lawsuit in 2023 with a tutoring company. This company automatically rejected anyone over the age of forty- its users were younger students who turned up complaining about broken microphones for online classes in college webcam setups and incomplete registration forms.

B. Fairness Metrics and Mitigation

Key fairness metrics in recruitment AI include: (1) Demographic Parity—equal positive prediction rates across protected groups; (2) Equal Opportunity—equal True Positive Rates for qualified candidates from different groups; (3) Predictive Parity—equal

precision across groups; (4) Individual Fairness—similar individuals receive similar treatment; (5) Counterfactual Fairness—decisions remain unchanged if protected attributes were different [23].

Mitigation strategies span the ML pipeline: Pre-processing includes resampling, reweighting training data, and removing biased features. In-processing includes adversarial debiasing and fairness constraints in loss functions. Post-processing includes threshold adjustment and re-ranking for diverse recommendations. Regulatory frameworks such as NYC Local Law 144 require bias audits for automated employment decision tools, and the EU AI Act will regulate high-risk AI systems including recruitment tools [24].

VIII. EVALUATION METHODOLOGY

A. ML Performance Metrics

ML performance evaluation metrics for job recommendation include: Precision@K and Recall@K measure accuracy and coverage of top-K recommendations; NDCG (Normalized Discounted Cumulative Gain) evaluates ranking quality with higher weights for top positions; MRR (Mean Reciprocal Rank) measures the average position of the first relevant item; Hit Rate@K measures the proportion of having at least one relevant item in top-K; AUC-ROC for binary classification tasks such as will-apply prediction [25].

B. XAI Quality Assessment

XAI quality evaluation spans multiple dimensions: (1) Fidelity—explanation accurately reflects actual model behavior; (2) Comprehensibility—users can understand the explanation; (3) Sufficiency—explanation provides enough information to understand the decision; (4) Stability—explanations are consistent for similar inputs; (5) Actionability—explanation guides specific actions. User studies with A/B testing measure trust, satisfaction, and decision quality with/without explanations [26].

TABLE III
EVALUATION METRICS FOR ML-XAI CAREER GUIDANCE SYSTEMS

Metric	Type	Formula/Description	Interpretation	Typical Values
Precision@K	ML	Relevant in top-K / K	Top-K accuracy	0.15-0.35 (K=10)
NDCG@K	ML	DCG@K / IDCG@K	Ranking quality	0.65-0.85
MRR	ML	Mean(1/rank of first relevant)	First relevant item position	0.55-0.75
Fidelity	XAI	Correlation with model behavior	Explanation accuracy	0.85-0.95
User Trust	XAI	Likert scale survey (1-7)	User confidence	5.2-6.1 (with XAI)
Demographic Parity	Fairness	$\|P(\hat{Y}=1\|A=0) - P(\hat{Y}=1\|A=1)\|$	Cross-group fairness	< 0.10 desired

IX. DISCUSSION

A. Current Challenges

Cold-start problem: Lack of data for new users or emerging occupations remains a significant challenge. Knowledge transfer and zero-shot learning are research directions to address this issue. Data quality and standardization: Career data is often unstandardized, with the same skill described in multiple ways. Skill taxonomies such as ESCO and O*NET need continuous updates to keep pace with labor market changes [27].

Accuracy-Interpretability trade-off: High-accuracy models (deep neural networks) are often difficult to explain, while easily interpretable models (decision trees) have lower accuracy. Research on interpretable-by-design models attempts to bridge this gap. Evaluation standardization: Lack of standardized benchmark datasets and evaluation protocols for career recommendation systems makes method comparison difficult [28].

B. Future Research Directions

Multi-modal learning: Integrating multiple data modalities including text (resume, job descriptions), images (portfolio, certificates), video (interview recordings), and behavioral data (click patterns, time spent). Multi-modal fusion has potential to provide comprehensive views of candidate-job fit. Causal

inference: Shifting from correlation-based to causation-based recommendations to understand "why" someone succeeds in a specific career. Causal ML can identify effective interventions in career development [29].

Real-time adaptation: Systems need to adapt quickly to changing market conditions and user preferences. Online learning and continual learning techniques allow models to update continuously. Lifelong career guidance: Expanding from one-time job matching to continuous career development support throughout working life, with periodic re-assessment and updated recommendations. Human-AI collaboration: Designing systems to augment human career counselors rather than replace them, with AI providing data-driven insights and counselors providing empathy and contextual understanding [30].

X. CONCLUSION

This paper has presented a comprehensive systematic review of Hybrid ML-XAI architectures in intelligent career guidance systems. Through analysis of 95+ research papers, we identified five major trends: Deep Learning hybrids, Knowledge Graphs, XAI integration, Reinforcement Learning, and Large Language Models. Results demonstrate that hybrid approaches significantly improve both accuracy and interpretability compared to single-method approaches.

The proposed five-layer integration framework (Data, Knowledge, ML, XAI, Presentation) provides a blueprint for developing next-generation career guidance systems. Fairness and privacy issues need to be addressed through bias mitigation strategies and federated learning. Future development directions include multi-modal learning, causal inference, and lifelong career guidance with human-AI collaboration.

This study contributes to the literature by: (1) providing a comprehensive taxonomy of ML and XAI methods in career guidance; (2) systematically comparing approaches with evaluation metrics; (3) identifying research gaps and future directions; (4) proposing an integration framework to guide future implementations. The findings have practical implications for researchers, practitioners, and policymakers in career guidance and educational technology domains.

Acknowledgment

The authors would like to thank the Faculty of Information Technology and the leadership of Ha Long University, Quang Ninh; the Faculty of Signal Processing and Communication; the Faculty of Electronic Engineering 1, Posts and Telecommunications Institute of Technology (PTIT), Vietnam; CMC University, Hanoi, Vietnam; and Hanoi University of Industry for providing the conditions for the authors to complete this research.

REFERENCES

1. Y. Zhang et al., "Educational data mining techniques for student performance prediction: Method review and comparison analysis," Frontiers in Psychology, vol. 12, p. 698490, 2021.
2. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. ACM SIGKDD, 2016, pp. 1135-1144.
3. S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.
4. T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artificial Intelligence, vol. 267, pp. 1-38, 2019.
5. C. Qin et al., "An enhanced neural network approach to person-job fit in talent recruitment," ACM Trans. Inf. Syst., vol. 38, no. 2, pp. 1-33, 2020.
6. S. Liu et al., "Hybrid job recommendation using deep learning and collaborative filtering," Knowledge-Based Systems, vol. 215, p. 106735, 2021.
7. H. Sun et al., "Career trajectory prediction with LSTM networks," Expert Systems with Applications, vol. 189, p. 116089, 2022.
8. J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language

understanding," in Proc. NAACL-HLT, 2019, pp. 4171-4186.
9. W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in Advances in Neural Information Processing Systems, vol. 30, 2017.
10. Y. Zhou et al., "Reinforcement learning for personalized learning path recommendation," IEEE Trans. Learn. Technol., vol. 16, no. 2, pp. 189-203, 2023.
11. F. Zhang et al., "Explainable job recommendation with LIME," in Proc. RecSys, 2022, pp. 234-243.
12. S. Jain and B. C. Wallace, "Attention is not explanation," in Proc. NAACL-HLT, 2019, pp. 3543-3556.
13. R. Guidotti et al., "A survey of methods for explaining black box models," ACM Computing Surveys, vol. 51, no. 5, pp. 1-42, 2018.
14. R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in Proc. FAT*, 2020, pp. 607-617.
15. A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82-115, 2020.
16. LinkedIn Engineering, "Building the LinkedIn Knowledge Graph," LinkedIn Engineering Blog, 2023.
17. European Commission, "ESCO: European Skills, Competences, Qualifications and Occupations," 2024. [Online]. Available: https://esco.ec.europa.eu/
18. T. Eloundou et al., "GPTs are GPTs: An early look at the labor market impact potential of large language models," arXiv preprint arXiv:2303.10130, 2023.
19. J. Li et al., "GIRL: Generative job recommendations with large language model," arXiv preprint arXiv:2307.02157, 2023.
20. D. Yao et al., "FedRKG: A privacy-preserving federated recommendation framework via knowledge graph enhancement," arXiv preprint arXiv:2401.11089, 2024.
21. J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," Reuters, Oct. 2018.
22. K. Wilson and A. Caliskan, "AI tools show biases in ranking job applicants' names according to perceived race and gender," in Proc. AAAI/ACM AIES, 2024.
23. N. Mehrabi et al., "A survey on bias and fairness in machine learning," ACM Computing Surveys, vol. 54, no. 6, pp. 1-35, 2021.
24. European Parliament, "EU Artificial Intelligence Act," 2024.
25. X. He et al., "Neural collaborative filtering," in Proc. WWW, 2017, pp. 173-182.
26. F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.
27. S. Ji et al., "A survey on knowledge graphs: Representation, acquisition, and applications," IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 2, pp. 494-514, 2021.
28. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206-215, 2019.
29. J. Pearl and D. Mackenzie, The Book of Why: The New Science of Cause and Effect. Basic Books, 2018.
30. D. Gunning et al., "XAI—Explainable artificial intelligence," Science Robotics, vol. 4, no. 37, p. eaay7120, 2019.