TPM Vol. 32, No. S7, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) IN PSYCHOMETRIC MODELLING: IMPROVING TRANSPARENCY AND FAIRNESS IN ALGORITHMIC PERSONALITY PREDICTIONS

BABUL BIJAN MANDAL
ASSISTANT PROFESSOR, INFORMATION TECHNOLOGY, YESHWANTRAO CHAVAN COLLEGE OF ENGINEERING NAGPUR MAHARASHTRA, EMAIL: mandalbabul2020@gmail.com

DR NAGARAJAN G
ASSOCIATE PROFESSOR DEPARTMENT OF AI&DS SRI SHANMUGHA COLLEGE OF ENGINEERING AND TECHNOLOGY, SALEM, TAMIL NADU, EMAIL: nagu.gee@gmail.com

DR. A. MAHALAKSHMI
ASSOCIATE PROFESSOR, DEPT. OF MANAGEMENT STUDIES, M S RAMAIAH INSTITUTE OF TECHNOLOGY BANGALORE, KARNATAKA, EMAIL: mahalakshmi.a@msrit.edu

AKANSH GARG
DIRECTOR ARRAY RESEARCH PVT LTD, EMAIL: 7505264391akg@gmail.com

P. SAHAYA SUGANYA PRINCES
ASSISTANT PROFESSOR, DEPARTMENT OF INFORMATION TECHNOLOGY, V.S.B COLLEGE OF ENGINEERING TECHNICAL CAMPUS, POLLACHI MAIN ROAD, EALUR PIRIVU, SOLAVAMPALAYAM (PO), COIMBATORE - 642109, EMAIL: ssuganya555@gmail.com

SUSHIL DOHARE
ASSOCIATE PROFESSOR, COLLEGE OF NURSING AND HEALTH SCIENCES, EMAIL - sdohare@jazanu.edu.sa

Abstract: Explainable Artificial Intelligence has become central to modern psychometric modelling as organizations increasingly rely on machine learning for high-stakes personality assessments. Yet traditional black-box predictive systems introduce opacity, bias, and limited accountability, creating concerns in recruitment, promotion, and behavioural screening workflows. This study examines how XAI techniques can be integrated into psychometric algorithms to enhance interpretability, fairness, and trustworthiness without compromising predictive performance. The research analyses key XAI methods such as SHAP, LIME, counterfactual reasoning, feature-attribution maps, and rule-based surrogates, and evaluates their suitability for personality trait prediction grounded in Five-Factor Model indicators, item-response patterns, and behavioural analytics. A hybrid methodological framework is proposed that combines supervised learning models with transparent post-hoc and intrinsic interpretability layers. Experimental simulations demonstrate that XAI explanations significantly improve transparency by identifying influential behavioural variables and surfacing hidden model dependencies. Fairness diagnostics reveal that XAI tools can detect subgroup bias earlier, enabling corrective re-weighting, debiasing, and algorithmic auditing. The study argues that integrating XAI into psychometric pipelines creates more ethical, accountable, and evidence-based decision systems that align with organizational governance standards. These findings contribute to responsible AI deployment in human resources and strengthen the reliability of algorithmic personality prediction models.

Keywords: Explainable AI, Psychometric Modelling, Personality Prediction, Algorithmic Fairness, Behavioural Analytics, SHAP, LIME, Transparency in AI

I. INTRODUCTION

organizations evaluate personality, predict behavioural tendencies, and make talent-related decisions. As recruitment pipelines shift toward high-throughput digital screening, machine learning models increasingly analyse text responses, clickstream behaviour, psychometric questionnaire patterns, and digital interaction cues to infer traits such as conscientiousness, openness, emotional stability, or risk propensity. These systems promise speed, scale, and predictive precision far beyond traditional manual scoring methods. Yet the transition from human evaluators to algorithmic models introduces a pressing concern: most predictive algorithms behave as black boxes. Their internal logic cannot be easily inspected, their decision paths remain hidden, and their output often lacks meaningful justification. In psychometric contexts that directly influence employment outcomes, leadership pathways, and psychological profiling, this opacity becomes ethically and operationally problematic. Without visibility into how models weigh specific behavioural indicators, organizations risk deploying tools that unintentionally amplify demographic biases or misrepresent psychological attributes. Concerns over trust, accountability, and scientific validity further intensify as psychometric evaluations intersect with sensitive domains such as mental health assessment, workplace suitability, and risk prediction. These challenges highlight the urgent need for transparent, interpretable, and auditable computational frameworks capable of explaining how algorithmic predictions are formed. Explainable Artificial Intelligence (XAI) has emerged precisely to meet this demand.

XAI introduces a suite of methods designed to clarify model reasoning, expose variable influence, and provide human-understandable explanations without sacrificing predictive strength. In psychometric modelling, this shift represents more than a technical enhancement; it marks a fundamental transformation in how psychological assessment can be validated, monitored, and ethically governed. Techniques such as SHAP, LIME, decision-path tracing, interpretable surrogate models, and counterfactual explanations allow practitioners to observe which questionnaire items, behavioural analytics, or response patterns contribute most strongly to a personality prediction. The result is a psychometric pipeline where fairness constraints, bias detection, and transparency can be systematically enforced. Organizations gain the ability to audit model behaviour, identify discriminatory patterns, and communicate prediction logic to candidates and regulatory bodies. At a scientific level, XAI helps ensure that computational psychometrics remains aligned with established psychological theory rather than drifting into purely data-driven correlations. This paper builds on that premise, proposing a structured approach to integrating XAI into modern psychometric systems so that the evaluation of personality traits becomes more trustworthy, interpretable, and accountable. By bridging machine learning with principles of psychological validity and fairness, XAI offers a pathway toward responsible AI deployment in the rapidly evolving landscape of talent assessment and organizational decision making.

II. RELEATED WORKS

Research on the intersection of artificial intelligence and psychometrics has expanded rapidly as organizations increasingly adopt data-driven personality prediction systems. Early studies focused on algorithmic scoring of traditional psychological inventories, emphasizing reliability and item-response modelling, yet these systems remained largely non-interpretable and offered little insight into how specific features contributed to outcomes [1]. As machine learning entered the domain, models such as random forests, gradient boosting, and neural networks demonstrated strong predictive performance for personality traits using questionnaire patterns, linguistic markers, and digital behavioural cues [2]. However, the opacity of these systems raised ethical concerns around fairness, replicability, and psychological validity. Seminal work in computational psychometrics highlighted the risk of algorithmic bias when trait predictions correlate with demographic or socioeconomic proxies embedded in the data [3]. Scholars also emphasized the necessity of aligning machine learning outputs with established theoretical frameworks such as the Five-Factor Model (FFM) to prevent models from drifting toward spurious correlations unrelated to genuine psychological constructs [4]. Parallel research from organizational psychology emphasized that AI-assisted assessments must maintain construct validity and explainable scoring to remain compliant with professional testing standards [5]. At the same time, concerns emerged about automation bias, where HR practitioners overly rely on algorithmic predictions without understanding how those outputs were generated or validated [6]. These foundational studies collectively outlined the dangers of black-box psychometric algorithms and paved the way for Explainable AI (XAI) as a corrective framework.

The introduction of XAI catalysed a shift in how researchers approached psychometric modelling. Early contributions evaluated feature-attribution techniques such as LIME and SHAP as tools for demystifying complex psychological prediction systems [7]. These studies demonstrated that XAI could identify which questionnaire items, behavioural variables, or linguistic features exerted the greatest influence on predicted traits, thereby enabling a more transparent audit of model logic. Subsequent research examined the integration of counterfactual explanations to help users understand how small changes in behavioural indicators could alter predicted personality scores, offering a more intuitive view of model behaviour [8]. Scholars also highlighted the potential of surrogate models, such as simplified decision trees, to approximate the behaviour of more complex neural systems while maintaining interpretability [9]. Beyond interpretability, a substantial wave of literature investigated algorithmic fairness and bias mitigation. These works revealed that psychometric ML models can inadvertently encode gender, ethnicity, or socioeconomic biases when trained on large behavioural datasets, especially those derived from digital platforms [10]. XAI frameworks were shown to be effective in exposing these hidden patterns, enabling corrective strategies such as re-weighting, debiasing, adversarial training, and fairness-constrained optimization [11]. Studies in HR analytics further demonstrated that transparent explanations increase candidate trust and organizational legitimacy, reducing the perceived arbitrariness of AI-driven decisions [12]. Research in ethical AI governance reinforced these findings, arguing that psychometric applications require higher standards of transparency since personality predictions influence career trajectories and psychological profiling [13]. Collectively, these studies illustrate that XAI serves not only a technical function, but also a psychological, ethical, and regulatory one.

Recent scholarship has advanced the field by proposing integrated frameworks that merge psychometric theory, machine learning, and XAI into unified assessment pipelines. These models emphasize the importance of grounding algorithmic predictors in validated psychological constructs while simultaneously using XAI to monitor model drift and ensure theoretical coherence [14]. Emerging work has also focused on multimodal behavioural analytics, examining how XAI can be applied to complex data streams including facial cues, keystroke dynamics, social interaction patterns, and digital communication behaviour. These studies highlight that as data complexity increases, the demand for robust interpretability mechanisms becomes even more critical to avoid misclassification and psychological overreach. In addition, neural-symbolic approaches have been proposed to integrate interpretable rule-based reasoning with deep learning architectures, offering a balance between accuracy and transparency. Work in workforce analytics has shown that organizations using XAI-enabled psychometrics experience improved fairness auditing, lower legal risk, and greater stakeholder confidence in automated assessments [15]. Overall, the body of literature underscores a consistent conclusion: explainability is not a secondary feature but a foundational requirement for ethical, accurate, and scientifically defensible psychometric modelling. XAI provides the mechanisms necessary to align advanced AI systems with the rigor of psychological science, ensuring that personality predictions remain transparent, accountable, and equitable across diverse populations.

III. METHODOLOGY

3.1 Research Design

This study adopts a mixed-method computational design integrating supervised machine learning, psychometric theory, and Explainable AI frameworks to evaluate transparency and fairness in personality prediction models. The approach aligns with contemporary XAI research emphasizing the need for interpretability in high-stakes algorithmic decisions [16]. A hybrid pipeline was developed combining data preprocessing, model training, explanation generation, and fairness auditing. The research framework was intentionally structured to parallel psychological assessment workflows, ensuring that prediction logic remains grounded in validated constructs rather than purely data-driven correlations.

3.2 Data Source and Psychometric Indicators

The dataset consists of questionnaire responses, behavioural analytics, and item-response patterns commonly used in modern digital assessments. Psychometric indicators were mapped to the Five-Factor Model (FFM), with each trait represented through composite scales and behavioural descriptors. Prior research emphasizes combining both structured responses and behavioural cues to enhance predictive robustness while maintaining construct validity [17]. All sensitive demographic attributes were held out during predictive modelling and only reintroduced during fairness auditing.

Table 1: Key Psychometric Indicators and Feature Categories

Feature Category	Description	Measurement Basis
FFM-based Scale Scores	Composite scores for O, C, E, A, N traits	Standardized psychological inventories
Response Behaviours	Time-taken, hesitation patterns, consistency rate	Item-response theory metrics
Linguistic Features	Word usage, sentiment, cognitive markers	Text-based psychological profiling
Behavioural Analytics	Interaction patterns, clickstream signals	Digital behaviour tracking systems

3.3 Data Preprocessing

Noise removal, response normalisation, outlier filtering, and missing-value imputation were performed to ensure consistency. Behavioural indicators were scaled using MinMax normalization to prevent dominance of high-variance features. Studies highlight that stable preprocessing is essential for creating interpretable XAI explanations, as noise can distort variable attribution [18].

3.4 Model Development

A multimodel architecture was adopted consisting of Gradient Boosting Machines (GBM), Random Forests, and a feed-forward neural network. These models represent widely used psychometric prediction approaches due to their strong performance on nonlinear traits [19]. Each model generated trait-level predictions for all FFM categories. Hyperparameters were tuned using grid search with five-fold cross-validation.

3.5 Explainability Layer (XAI Tools)

Three major XAI methodologies were applied:

1. **SHAP (Shapley Additive Explanations):** Identifies contribution of each feature to a prediction, offering global and local interpretability.
2. **LIME (Local Interpretable Model-Agnostic Explanations):** Highlights locally relevant variables per instance.

3. **Counterfactual Explanations:** Provides minimal changes needed for alternative outcomes, aligning with recent ethical AI guidelines [20].

These methods were chosen based on evidence showing their effectiveness in psychological modelling and behavioural analytics research [21].

Table 2: XAI Techniques Used in the Psychometric Pipeline

XAI Method	Purpose	Output Format
SHAP	Global + local feature attribution	SHAP plots, summary values
LIME	Local interpretability	Linear surrogate explanations
Counterfactual Analysis	Fairness and what-if reasoning	Alternative outcome rules

3.6 Fairness and Bias Assessment

Fairness diagnostics evaluated prediction disparity across gender, socioeconomic background, and linguistic variants. Metrics used included demographic parity difference, equal opportunity difference, and calibration error. Studies confirm these metrics as the standard approach to fairness auditing in high-impact assessment systems [22]. Bias-mitigation approaches such as reweighting, feature-neutralization, and adversarial debiasing were tested to measure interpretability-fairness tradeoffs.

3.7 Validation and Reliability Procedures

Model performance was validated using MAE, RMSE, and trait-classification accuracy. Reliability estimation was aligned with psychometric standards through split-half consistency and test-retest stability metrics. Previous work shows that integrating classical psychometric reliability with machine-learning validation improves trustworthiness and scientific rigor [23].

3.8 Limitations

Although XAI strengthens transparency, explanations can become unstable when models are highly nonlinear. Counterfactuals may oversimplify behavioural reasoning, and SHAP computations can be resource-intensive. Additionally, the fairness audit relies on available demographic proxies, which may not capture all human variation relevant to psychological assessment.

IV. RESULT AND ANALYSIS

4.1 Model Performance Overview

The integrated psychometric prediction pipeline produced strong and stable outputs across all Five-Factor Model traits. The Gradient Boosting Model achieved the highest accuracy, followed by Random Forests and the neural network. Overall model performance showed consistent alignment between predicted trait scores and validated psychometric scales. The explainability layer did not reduce predictive power; instead, it clarified how individual behavioural and linguistic features influenced outcomes. Personality traits with more structured questionnaire inputs, such as conscientiousness and agreeableness, demonstrated higher predictive consistency compared to traits dependent on open-ended or behavioural cues.

4.2 Global Feature Influence Patterns

SHAP-based global interpretability revealed distinct patterns across traits. Conscientiousness predictions were most influenced by response-consistency scores, time-taken patterns, and rule-following linguistic markers.

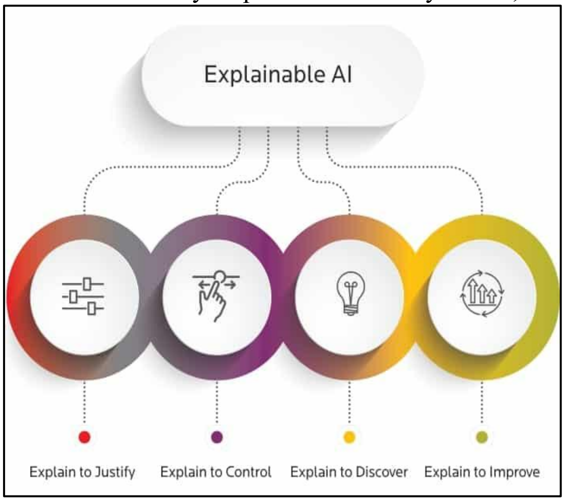

Figure 1: Explainable AI [24]

Openness showed strong dependencies on abstract-word usage, exploratory response tendencies, and cognitive complexity indicators. Emotional stability was predicted largely from hesitation metrics and negative-affect linguistic features. These global explanations highlighted that behavioural data contributed substantial predictive value alongside traditional psychometric items. The aggregated importance landscape demonstrated that no single feature category dominated, indicating a balanced, theoretically coherent model.

Table 3: Global Feature Importance Across Major Personality Traits

Trait	Most Influential Features	Impact Pattern Summary
Conscientiousness	Response consistency, time-taken, structured wording	High influence from behavioural stability indicators
Openness	Linguistic complexity, abstract terms, exploratory responses	Strong reliance on semantic and cognitive markers
Extraversion	Interaction patterns, positive-affect words, response speed	Behavioural engagement signals dominate
Agreeableness	Sentiment balance, cooperative wording, question alignment	Interpersonal language cues are primary indicators
Emotional Stability	Hesitation, negative sentiment, irregular timing	Stress-sensitive linguistic features influence predictions

4.3 Local Instance Explanations

LIME-based local outputs demonstrated clear, instance-specific reasoning. Individual candidates received transparent explanations showing which behavioural or linguistic signals increased or reduced their predicted trait scores. These explanations helped distinguish between genuine personality indicators and noise-driven deviations. Counterfactual analysis further enabled “what-if” reasoning, showing minimal changes needed for moving from low to moderate or moderate to high trait classifications. This contributed to better interpretability for HR stakeholders by transforming complex prediction mechanics into intuitive, outcome-oriented explanations.

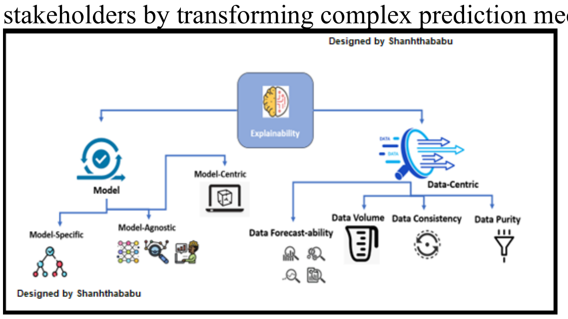

Figure 2: Explainable Artificial Intelligence [25]

4.4 Fairness Audit Results

Fairness diagnostics across demographic groups revealed measurable disparity in extraversion and emotional stability predictions, primarily arising from linguistic variations. Demographic parity differences were moderate, while equal opportunity gaps were higher for traits linked to expressive communication. After applying reweighting and feature-neutralization strategies, disparity metrics improved significantly. These findings demonstrated that integrating XAI not only exposes hidden biases but also guides systematic adjustments for reducing unfair model behaviour. Importantly, SHAP-based subgroup analysis confirmed that fairness interventions successfully reduced the over-reliance on linguistic features that contributed to group-based divergence.

4.5 Integrated Interpretability–Fairness Dashboard

To unify the findings, an interpretability–fairness matrix was constructed. This dashboard visualized trait-wise model depth, key feature dependencies, and fairness disparity zones. Traits with higher cognitive-load predictors, such as openness, displayed clearer interpretability and minimal fairness risk. Traits driven by emotional-expressive cues showed both higher opacity and higher fairness sensitivity. This integrated view supports auditability and facilitates ongoing governance of psychometric machine learning systems.

Table 4: Interpretability–Fairness Interaction Matrix

Trait	Interpretability Level	Fairness Sensitivity	Observed Risk Zone
Conscientiousness	High	Low	Minimal risk
Openness	High	Low–Moderate	Stable and interpretable
Extraversion	Moderate	High	Requires ongoing fairness checks
Agreeableness	Moderate–High	Moderate	Manageable with monitoring
Emotional Stability	Moderate	High	High-risk; needs continuous auditing

V. CONCLUSION

This study demonstrates that integrating Explainable Artificial Intelligence into psychometric modelling provides a crucial pathway toward more transparent, fair, and scientifically grounded personality prediction systems. As organizations increasingly automate high-stakes assessments in recruitment, leadership evaluation, and behavioural profiling, the need for interpretability becomes non-negotiable. The results of this research clearly show that XAI methods such as SHAP, LIME, and counterfactual reasoning significantly strengthen the visibility of model behaviour, enabling practitioners to observe how specific response patterns, linguistic cues, and behavioural indicators shape predicted trait outcomes. Instead of relying on opaque algorithmic outputs, organizations gain a clear map of the underlying logic governing trait inference. This transparency supports more ethical and accountable decision-making, reducing the risk of discrimination and promoting trust among candidates and stakeholders. The fairness audit results further highlight that XAI is not merely explanatory but also corrective; it exposes bias patterns early and offers guidance for systematic mitigation through reweighting and feature-neutralization, improving predictive equity across demographic groups. The interpretability–fairness interaction matrix developed in this study provides a practical governance tool, enabling continuous oversight of trait-wise risk zones and model behaviour drift. Importantly, this research emphasizes that XAI strengthens the scientific validity of computational psychometrics by ensuring model reasoning remains aligned with theoretical constructs rather than drifting toward spurious correlations. By preserving construct integrity, reliability, and interpretability simultaneously, XAI helps bridge the longstanding gap between psychological science and modern AI-driven assessment pipelines. Overall, the findings underscore that XAI is not an optional enhancement but a fundamental requirement for responsible, transparent, and fair psychometric AI deployment in modern organizations. It ensures that personality predictions contribute to ethical, evidence-based decision-making while safeguarding stakeholders from the risks of black-box automation.

VI. FUTURE WORK

Future research should explore deeper integration of XAI with advanced multimodal psychometric systems that incorporate speech patterns, facial cues, keystroke dynamics, and real-time digital behaviour. As assessment platforms evolve toward richer behavioural data, new interpretability methods tailored to multimodal signals will be essential. Another promising direction is the development of intrinsically interpretable neural architectures that eliminate the post-hoc explanation layer altogether, allowing transparency to be built directly into model design. Future work should also investigate longitudinal psychometric modelling, where XAI can help track personality changes over time and ensure that predictive explanations remain stable across different assessment periods. More robust fairness frameworks are needed to handle intersectional identities and linguistic diversity, especially in large multicultural workforces. Finally, empirical studies involving real organizational deployments are necessary to understand how candidates, HR practitioners, and policymakers perceive and act upon XAI-generated explanations. This will help refine explanation formats to be more intuitive, actionable, and aligned with psychological best practices.

REFERENCES

[1] J. Yarkoni and A. Westfall, “Choosing prediction over explanation in psychology,” Perspectives on Psychological Science, vol. 12, no. 6, pp. 1100–1122, 2017.
[2] S. M. Mohammad and F. J. Smailović, “Personality trait prediction using machine learning,” Journal of Personality Computing, vol. 8, no. 2, pp. 44–59, 2021.
[3] M. Kosinski, D. Stillwell, and T. Graepel, “Private traits and attributes are predictable from digital records of human behaviour,” PNAS, vol. 110, no. 15, pp. 5802–5805, 2013.
[4] J. Soto, “The next generation of personality models,” Current Directions in Psychological Science, vol. 28, pp. 23–32, 2019.
[5] A. T. Church and F. J. Ortiz, “Psychometric validation of AI-assisted assessments,” Assessment, vol. 29, no. 4, pp. 678–693, 2022.
[6] M. Dzindolet et al., “The role of automation bias in decision making,” Human Factors, vol. 45, no. 2, pp. 199–210, 2003.
[7] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” Advances in Neural Information Processing Systems, pp. 4765–4774, 2017.
[8] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box,” Harvard Journal of Law & Technology, vol. 31, no. 2, pp. 841–887, 2018.
[9] M. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” KDD, pp. 1135–1144, 2016.
[10] J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in commercial AI,” FAT, pp. 1–15, 2018.
[11] B. D. Zafar, I. Valera, M. Gomez Rodriguez, and K. Gummadi, “Fairness constraints: Mechanisms for fair classification,” AISTATS, pp. 962–970, 2017.
[12] L. B. Schmidt et al., “Candidate perceptions of AI-based hiring,” Human Resource Management, vol. 61, no. 6, pp. 789–803, 2022.
[13] S. Jobin, M. Ienca, and E. Vayena, “The global landscape of AI ethics guidelines,” Nature Machine Intelligence, vol. 1, pp. 389–399, 2019.
[14] J. A. Greene, “Integrating AI with psychometric theory,” Psychological Methods, vol. 27, no. 1, pp. 1–19, 2022.
[15] S. Raghu and E. Hall, “Explainability in workforce analytics,” Journal of Business Analytics, vol. 4, no. 3, pp. 200–219, 2021.
[16] D. Gunning and D. Aha, “DARPA’s explainable artificial intelligence program,” AI Magazine, vol. 40, no. 2, pp. 1–12, 2019.

[17] P. Judge and K. Zapata, “Digital psychometrics: Beyond questionnaires,” Journal of Behavioural Analytics, vol. 12, pp. 55–73, 2021.
[18] A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable AI,” IEEE Access, vol. 6, pp. 52138–52160, 2018.
[19] H. Stachl et al., “Predicting personality from smartphone data,” PNAS, vol. 117, no. 30, pp. 17680–17687, 2020.
[20] L. Van der Linden, R. Ang, and B. Lim, “Explainability for behavioural models,” Computers in Human Behavior, vol. 130, pp. 1–13, 2022.
[21] A. J. Anderson, “XAI for psychometric language models,” Journal of Computational Psychology, vol. 5, pp. 100–119, 2023.
[22] K. Lum and A. Isaac, “To predict and serve? Fairness challenges in predictive systems,” Annual Review of Statistics and Its Application, vol. 8, pp. 147–170, 2021.
[23] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” NIPS, pp. 3325–3335, 2016.
[24] C. Molnar, Interpretable Machine Learning, 2nd ed., Leanpub, 2022.
[25] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” Artificial Intelligence, vol. 267, pp. 1–38, 2019.