J. Res. Technol. Eng. 6 (3), 2025, 303-312

Journal Of Research Technology & Engineering
www.jrte.org

Multi-modal Personality Predictions and Fairness aware AI Techniques: A Review of Methods, Bias Mitigation and Future Perspectives

M.P. Kandage, H.K.I.S. Lakmal

Faculty of Engineering, NSBM Green University, Sri Lanka
mpkandage@students.nsbm.ac.lk, isuru.l@nsbm.ac.lk

Received:28 June 2025; Revised: 30 June 2025; Accepted: 06 July 2025; Available online: 10 July 2025

Abstract: Multi-modal artificial intelligence (AI) in personality forecasting can have revolutionary effects in hiring, psychological diagnosis, and anthropomorphic interaction. Deep learning architectures based on audio, visual, and textual data via behavioral cues tend to be efficient when the correct hypothesis is stated in the context of the Five-Factor Model (OCEAN)or by deciphering benchmarks (e.g., the approach of the ChaLearn First Impressions V2 data). Nevertheless, such systems will threaten to heighten demographic biases (e.g., gender, ethnicity) due to the lack of representation, subjective labeling, or incorrect spurious correlations of the AI. The review collates the recent developments of multi-modal fusion methods, fairness measures (e.g., Equal Opportunity, Statistical Parity), and bias reduction algorithms in preprocessing (data balancing), in-processing (adversarial debiasing), and postprocessing (weighted integration). We also draw on transparency, consent, and regulatory aspects, as well as address ethical issues, and establish key research gaps: dynamic fairness adaptation, cross-domain generalization, explainable AI, and context-aware fusion. In our analysis, we stress the importance of interdisciplinary interventions to build responsible, equitable personality AI systems that can be implemented in high-stakes areas.

Index Terms: Bias, Fairness, Multi-modal Personality Prediction, Personality traits

1 INTRODUCTION

Machine analysis of personality has become more central to AI decision-making and is applied in recruitment, education, screening of mental health, and individualized human-computer interaction. Informing on personality attributes using multi-modal behavioral cues (e.g., facial expressions, prosody, language usage) is usually modeled through the Five-Factor Model (FFM or OCEAN: Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism), and such systems thus have the potential to offer objective, scalable measures [1],[2]. With the development of deep models and high-quality data sources, such as the ChaLearn First Impressions V2 [1], [2] multi-modal networks have been developed, which combine visual, audio, and textual information by extracting multi-modal features to attain leading predictive performance.

Alongside this development, however, there are severe concerns. Demographic biases affect all personality prediction systems such that performance differences affect the marginalized groups (e.g., ethnic minorities and non-binary individuals) disproportionately. These biases are caused by systematically biased training data, subjectivity in measurements such as (cultural) stereotyping in crowd-sourced labels, and spurious correlations in algorithms requiring a training data misrepresentation between demographic characteristics and personality traits [3],[4],[5]. These shortcomings will lead to discrimination in situations that are high stakes - e.g., unassertive introverted candidates being mislabeled in an AI assisted hiring context [6]. Therefore, the idea of fairness-aware AI methods becomes a concern that could not be postponed to a

secondary dimension.

There are three major objectives considered in this review:

- Integrate architectural developments of multi-modal prediction of personality (Section 2) such as deep feature extraction (LSTMs, CNNs) or fusion techniques (attention-based, transformer-guided).
- Organize fairness measures (e.g., Equal Opportunity, Statistical Parity) and bias predisposition (Section 3), putting them into context about how they affect demographic subgroups.
- Comparison of bias mitigation methods (Section 4) among the preprocessing (via data augmentation), in-processing (via adversarial debiasing) and postprocessing (via threshold adjustment) paradigms.

also discuss practical examples and ethical questions (Section 5), putting a particular emphasis on the conflicts between efficiency and equity. Last, to promote responsible deployment, we specify emergent challenges dynamic fairness, explainability, regulatory alignment and interdisciplinary research directions (Sections 6,7) to encourage responsible deployment. This review will help to encourage the evolution of personality AI systems that will be more than precise by way of building technical rigor and ethical foresight together, such systems will also be responsible.

2 MULTI-MODAL LEARNING AND FEATURE REPRESENTATION

Multi-modal learning combines diverse types of data or inputs to create a more complete model of user behavior and personality, e.g., visual, audible, and textual data. Popular models of applied sciences in this area frequently utilize deep convolutional neural networks (CNNs), and examples include VGG16 [7], Efficient Net [8], and pretrained backbones, such as ResNet, VGGish, ELMo, and so on [9]. Such architectures are known to be critical in encoding modality-specific features in a high-fidelity manner. Models normally include Long Short-Term Memory (LSTM) systems to help them capture temporal dependencies and sequential patterns.

The preprocessing pipeline within the multi-modal systems is based on such powerful tools as FFmpeg [10] video decoding, Librosa [11] audio feature extraction, OpenCV [12] level by level analysis and face detection. The tools enable consistent feature extraction among modalities by simultaneous extraction in a consistent temporal and spatial manner.

Fusion strategies Multi-modal fusion strategies are usually grouped into feature-level, score-level, and decision-level fusion strategies. In feature-level fusion, these raw or intermediate feature representations are concatenated, whereas in score-level fusion, outputs in terms of scores of the models are averaged, summed weighted, or, in trainable attention, the scores of the models. The decision-level fusion fuses the last estimates of the modality. Fusion methods that deal with attention have shown to perform better, since they now dynamically weight the modalities rather than fix it globally, making the model interpretable, and more generalizable [13], [14], [9]. Transformer-based cross-modal attention layers and graph convolutional networks to learn the complex inter-modal relationships and semantics alignment to capture are also investigated recently [15], [16].

3 FAIRNESS AND BIAS IN MULTI-MODAL PERSONALITY PREDICTION

Since the AI systems are becoming frequently involved in human-related behaviors like personality evaluation, fairness becomes a crucial part. The personality inference is a sensitive human property the inference of which, (as compared to generic prediction) is easily swayed by the demographic factors in terms of tone, expression, and behavior. This predisposes such systems to systematic discrimination against

the subgroup members, and groups in general, particularly when such a system is leveraged on a socially meaningful context, such as recruiting or psychiatric diagnostics. It is important to be aware of the causes of bias and the measurements resorted to in determining the level of fairness to create ethical and accountable AI systems.

3.1 Types of Bias

There are multiple sources of bias in the personality prediction systems, and they may put unclean marks on the results differently:

- Representation Bias: This occurs when training data will underrepresents some groups (e.g., ethnic minorities or non-dominant gender identities). It results in the lack of effective generalization with regards to such groups [3].
- Measurement Bias: Occurs when labels (e.g., OCEAN trait annotations) are imperfect in the sense that, they are arrested by subjective opinions or diverse labeling mechanisms, predominantly in crowd-sourced databases [4].
- Algorithmic Bias: Occurs when models are trained to learn spurious correlations, or even when they over-fit to the prevalent demographic trends. As an example, models can assign more visual attributes such as smiling to be more closely linked to extroversion in one gender than in the other enhancing any bias that exists in society [17], [5].

Since the subject of personality traits is behavioral and visually subtle, it is especially vulnerable to such bias. As an example, examples of affect or intent may not be well perceived in the models, and have lower performance, depending on culture cultural expressiveness or audio pitches.

3.2 Fairness Metrics

To evaluate the fairness of the predictive models, particularly those in personality AI, several measures group-based metrics have been proposed. The latter can assist in measuring inequalities between sensitive groups like gender and ethnicity:

- Equal Opportunity (EO) examines the quality of models that provide qualified members of all groups (e.g., different genders or ethnicities) with an equal opportunity of being successfully chosen. It is designed to compare the true positive rates of the same, or how frequently does the model predict the correct positive case. Proportional model under EO would be availed fairly by working the same way in every demographic with the ground truth on the positive.

This will make individuals of all groups equal in the chances of being hired appropriately [18].

- Statistical Parity (SP), sometimes termed demographic parity, quantifies the properties of the model to state whether the model is choosing members of disparate groups at similar rates, without regard to whether it is choosing them correctly. It points out scenarios where in an overall selection frequency more than other group of people, even though the decisions on correctness may not be assured as a model.
- Equal Accuracy (EA) checks the accuracy of the model in different groups. It examines the consistency by which the model can be trusted across all subgroups in both cases of correct positives and correct negatives. This is particularly critical in highly stakes application where stability of model performance is required to instill trust and fairness [17].
- In addition to group-based measures of fairness metric, such as Equal Opportunity, disparate mistreatment [19] assesses whether the distribution of classification errors (e.g., false positives/negatives) is equally distributed across groups that define demographic categories. This

plays an important role in predicting personality where such errors (e.g., categorizing introversion as neuroticism) can be very disadvantageous to some people.

All these measures can have a different conception of fairness, and they do not always correlate. Thus, in fairness-aware AI, one tends to consider the results of many metrics. In addition, the metrics can be computed separately depending on sensitive attributes such as gender or ethnicity to gain an idea of the fairness effects to each group. There has also been an idea to summarize fairness by taking average trait-wise differences across groups to help point out where disparities are most egregious (Shen et al., 2020) [17].

4 BIAS MITIGATION TECHNIQUES

The current approaches to bias vary and include strategies used by modern AI systems to mitigate bias in the machine learning pipeline. There are ways in which these approaches can be classified depending on their stage of intervention into model development.

4.1 Preprocessing Approaches

One major way of preprocessing is that which modifies the input data to minimize the bias before training. The most common method is data balancing, in which the sets of data are modified to equally reflect demographic, either through endless sampling of the underrepresented category of a set or oversampling of the oversampled category of a set [17], [20], [4]. Although this will help to minimize representational bias, the danger is overfitting or lack of generalizability particularly in cases where there is a small number of datasets or highly augmented using synthetic data sets [5], [18].

The other preprocessing approach is data augmentation, especially, this way is possible by the creation of the new samples by minority groups. The representation can be enhanced with synthetic data generation, e.g., using GAN (Generative Adversarial Networks) s or SMOTE (Synthetic Minority Over-sampling Technique), but one needs to be careful that the data remains realistic and does not introduce noise or distortion of the demographic characteristics [4], [21]. In one example, Zhao et al. implored the idea that, at the corpus level, limits can be used to curb gender bias amplification when producing data [21].

Preprocessing of this nature has been used in real life scenarios such as Pymetrics in balancing applicant information and avoiding recruitment biases [6]. On the same note, Retorio has also trained its video-based personality artificial intelligence on preprocessing pipelines with data cleansing on training data [22].

4.2 In-processing Approaches

In-processing methods bring in equality limits into the training process of a model. The most popular one is adversarial debiasing, which consists of training a predictor and an adversary jointly. The predictor wants to make correct choices and adversary wants to guess values of sensitive attributes based on the intermediate values of the model, making the predictor avoid paying attention to such sensitive attributes as the race or gender [4], [17], [5].

Fair representation learning is another approach in which the model learns embedding of data that are zero predictive of the target but have invariance with respect to a protected variable. These ideas were first discussed in the works by Zemel et al. in their Learning Fair Representations framework [20], and later bias amplification was solved by means of corpus-level constraints by such authors as Zhao et al. [21]. This is the concept that finds modern elaborations in donation-adaptation solutions and the separation between demographic and task-relevant characteristics [4], [18].

Zafar et al. [19] suggested using fairness constraints, which are incorporated in the optimization objective, where disparate mistreatment constraints have been introduced to guarantee an equal error rate among

demographic groups. This is practically relevant to subjects such as criminal justice risk assessment and finance where disparities in false positive rates may lead to discrimination in the system [4], [19].

4.3 Postprocessing Approaches

Postprocessing is an approach that is used once the model has been trained and can be particularly valuable when retraining is costly or otherwise daunting. An easy postprocessing scheme is using threshold adjustments among different groups to make performances identical metrics like the true positive rate [4], [18]. An example of bias elimination would be to lower the acceptability level of female candidates who attended personality based recruitment tests [6].

Weighted fusion in the post-hoc situation is a powerful method to use in multi-modal systems. This is actually single pooling of predictions made in different modalities (audio, video, text) with adjustable weights. In case of finding out that video inputs will introduce gender bias, their dependability can be reduced in decision-making [14], [17], [23]. Kaya et al. transfer this trivially to a video CV screening system, where performance is improved with less retraining [14].

Kampman et al. made a fascinating analysis of several of the fusion approaches that could be used to merge modalities without creating a bias, giving pointers on how such solutions can be realized practically [23]. These solutions have become more common in practical systems such as Retorio [22] where it is necessary to tune fairness to ensure ethical use of HR systems.

5 DATA SETS AND TOOLS

It is among the most established multi-modal personality prediction benchmarks, the ChaLearn First Impressions V2 dataset [1], [2], [14]. It is composed when 10,000 short video clips (the length is about 15 seconds) derived out of YouTube interviews are taken. Big Five personality traits are crowdsourced with each clip being coded in terms of these traits. The demographic metadata contained in this dataset, namely, gender and ethnicity, makes the data a very appropriate choice to analyze fairness and bias in AI systems [2], [17].

To streamline the experimentation, there are open-source tools with prebuilt pipelines like the Daniel Grimm GitHub repository [24]. These normally incorporate:

- Audio feature extraction (e.g., MFCCs) with librosa [11]
- Detection and tracking: OpenCV on frame level face detection [12],
- TensorFlow and PyTorch in the construction of deep neural networks [13].

There are a few more datasets and tools that widen the range of fairness-aware multi-modal personality prediction:

- Biel and Gatica-Perez [25], [26] are the YouTube Vlogs dataset in which they use naturalistic video blogs that are annotated with crowdsourced personality impressions. It works especially towards reviewing nonverbal behavior, like gestures and stare.
- Multi-modal inputs (audio/video, along with physiological markers in some cases) annotated with emotion, affect, demographic attributes are available in the AVEC (Audio/Visual Emotion Challenge) datasets [27]. Designed to recognize emotion, they are however applied increasingly in personalities prediction models as affective state rides hand in hand with personality traits.
- The PsyNet corpus (employed by Mairesse et al.) [28] integrates textual and acoustical signals of human communications under the controlled interviews. It is worth mentioning as it bridges the gap between personalities by classifying traits through the text-audio fusion investigation.

The pipelines of preprocessing are vital, and toolkits such as OpenFace [29] and OpenSMILE [30] are essential at this stage:

- Facial landmark-based detection, eye-gaze, head-pose tracking, etc. process can be performed in OpenFace, which allows high-resolution analysis of facial cues.
- Other language features that are presented by OpenSMILE are comprehensive acoustic features that include prosody features, spectral features, and voice quality features- these can be utilized in modeling emotions and also personality.

To make fairness optimization, there are tools like Fairlearn [14] and AIF360 (AI Fairness 360) [31] that provide Python modules to measure bias, compare models, and mitigation techniques. They are of particular use when it comes to operationalization of fairness at the levels of the preprocessing, in-processing, and postprocessing.

There are also synthetic balancing techniques such as SMOTE [16] to deal with data imbalance which is very useful when real-world situations that reflect true diversity is scarce.

Collectively, the above-mentioned datasets and tools present a pool of tools that can be used to develop transparent, reproducible, and fairness-aware AI models to build multi-modal personality prediction models.

6 APPLICATIONS

The desire to predict personality through the use of AI has a very significant potential to transform many different fields. Companies are even resorting to video CV and Ideal interview when recruiting to ensure that the recruitment process is simplified and that the human bias is avoided by evaluating the personality and soft skills of applicants prior to short listing them. These systems examine their non-verbal communications such as their facial expressions, voice tone and body language to give personality evaluations [14].

In mental health, these tools may assist in locating emotional emotions or acting abnormalities in texts or facial dynamics, providing non-medical assistance in foreseeable cases of psychological illnesses. This can be especially adapted in the telemedicine and online therapeutic scenario. Personality-aware systems in human-computer interaction (HCI) have the potential to respond, use a certain tone, or create an interface depending on the inferred personality of the user and as a result of their communication being more empathetic and personalized [25], [26].

Nonetheless, there are a number of ethical issues that go with incorporating the personality AI in sensitive applications. Issues of transparency and explainability shape up since the users do not know how these models render judgments as well as whether they are fairly applied. Informed consent is also a major concern, users might never suspect that their behavioral patterns or appearance might be used to make a conclusion about some of the most essential personal qualities. Moreover, without the proper regulatory review, particularly on the part of locales with liberal data protection laws, entry can be misused, including discriminatory filtering during employment or profiling in consumer data. Having an accountability man to blame in case biased or flawed results are achieved is also a problem that remains open in implementing these systems in a responsible way.

A number of real-world systems serve as evidence of the fact that AI-guided personality prediction has already found its application in sensitive fields, such as hiring and psychological profiling. Among the most high-profile ones, HireVue analyzes video interviews with the help of AI to assess the personality of candidates, their emotional expressiveness, and communication style based on facial expressiveness, tone

of the voice, and the language content. Bigger companies such as Unilever and Hilton have employed this system to facilitate the recruiting process in the purported quest to eliminate human bias. Nonetheless, the private sector has been subject to criticism by privacy watchdogs and researchers because of questions regarding how algorithms can be tested as being relatively transparent and harboring no discriminatory bias against any demographics [32]. Equally, Pymetrics sports neuroscience game-based profiling of tests and machine learning, and has been utilised into Fortune 500 recruitment pipelines [6]. Retorio is another sophisticated framework that applies Big Five (OCEAN) personality forecasts with multi-modal features of a video interview and can be used in hiring as well as customized coaching [22]. Such places embody an increasing tendency towards attempting to automate personality tests in critical situations with the aid of AI. Generative multi-modal models (e.g., GPT-4o) amplify stereotypic correlations (e.g., linking 'assertiveness' with male-presenting vocal tones) through RLHF (Reinforcement Learning from Human Feedback) training [3],[16]. Their implementation has raised ethical consideration over informed consent, algorithmic bias, and explainability as they provide efficiency and behavioral insights, but this contributes to the urgency of regulatory supervision and interdisciplinary governance

7 CHALLENGES AND FUTURE DIRECTIONS

Although interesting advances in multi-modal personality prediction have occurred, there are key issues that still need to be overcome to allow general-purpose, ethical and responsible use:

- Dynamic Fairness: Fairness is dynamic. Depending upon the task or situation of concern, dimensions like gender, ethnicity, age may or may not be very critical. The next generation model must be able to compensate fairness constraint set on a dynamically changing basis to correspond to the sensitiveness per context.
- Cross-Domain Generalization: The vast majority of models are only trained on small, domain specific data (such as YouTube interviews) and they may not extend to other domains (formal interviews, clinical setting, other cultures, etc). This restricts their applicability and also terms of bias transfer are questioned in different domains.
- Adaptive Fusion Strategies: Since fixed-weight fusion methods assume an equal weight of modalities (e.g., audio or video) in all samples, they are considered as not adaptive. In the practical contexts however, quality of data or relevance can be varied between the users. In the future, we must have adaptive fusion or sample-specific fusion which will learn to weight the possibilities best depending on confidence in the input or pertinence in the demography.
- Explainability and Interpretability: Some deep learning models are called black boxes, and it is hard to justify how decisions are reached. This transparency not only diminishes the trust of the users but also makes the auditing processes difficult. It becomes increasingly necessary to implement explainable AI (XAI) methods to give meaning to prediction rights.
- Regulatory Compliance and Data Ethics: Since personality prediction involves the processing of sensitive personal data, it is necessary to recognize the regulation as the General Data Protection Regulation (GDPR) and other frameworks of AI governance[33]. Model development must incorporate moral design principles, such as impartiality, responsibility and human involvement, fundamentally.

An interdisciplinary encounter must be used to deal with these issues, and it must not only involve the assistance of technical innovation, but also of the law, morals, psychology, and human-computer interaction fields. The proposed research area in the future is to develop the trust in AI systems that are fair, general,

and transparent in real-life settings.

8 Conclusion

Multi-modal personality prediction AI is at a crossroads: its promise to transform the recruitment process, diagnosing mental health, and communicating with computers is undeniable, but without high-level safeguards against fairness, integrating such tools into society will instead head towards reproducing existing inequities. The survey has unified architectural inquiry into feature combinations (e.g., attention mechanism, cross-modal transformer), quantified bias using metrics like Equal Opportunity, Statistical Parity and evaluated mitigation schemes across ML pipeline stages; data balance (preprocessing), adversarial debiasing (in-processing) and threshold correction (post processing). Still, there are some problems that cannot be forgotten:

- An overfitting to a domain may limit the possibilities to externally transfer it to datasets with higher scores (e.g., ChaLearn V2 interviews).
- The contextual modality relevance is not considered in static fusion strategies.
- Black-box decision-making negates the exigence of transparency in high-stakes applications.

Among priorities, three have become urgent and require interdisciplinary approaches to be addressed in the future:

- Creating dynamic fairness frameworks that may change fairness constraints in a case-by-case manner (e.g., a higher degree of parity in the hiring department in contrast to enjoyment).
- Transiting to explainable AI (XAI) approaches to audit the trait-inference logic to win the confidence of stakeholders.
- Constructing ethical-legal guard rails, e.g., a standardized methodology of consent and compliance tools to regulate (GDPR, AI Act).

To achieve equity in personality AI, algorithmic innovation is not enough; it must be accompanied by a long-term conversation process involving technologists, psychologists, policymakers, and the communities that have been affected. Such collaboration is the only way to make sure that these systems evolve to not only be technically accurate, but also socially responsible tools of human empowerment.

References

[1] S. Escalera, B. Chen, A. Places, M. Oliu, C. Corneanu, X. Baro, H. Jair Escalante, I. Guyon, and S. Escalera, "ChaLearn LAP 2016: First round challenge on first impressions – dataset and results," in Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW), 2016, pp. 459–473.

[2] H. J. Escalante, V. Ponce-López, J. Wan, M. A. Riegler, B. Chen, A. Clapés, S. Escalera, I. Guyon, X. Baró, P. Halvorsen, and M. Müller, "Modeling, recognizing, and explaining apparent personality from videos," IEEE Trans. Affective Comput., vol. 11, no. 1, pp. 2–20, 2020, doi: 10.1109/TAFFC.2017.2736198.

[3] B. Friedman and H. Nissenbaum, "Bias in computer systems," ACM Trans. Inf. Syst., vol. 14, no. 3, pp. 330–347, 1996.

[4] S. Jia, L. Wang, and M. Xu, "Fairness-aware machine learning for predictive modeling: A survey," ACM Comput. Surv., vol. 55, no. 1, pp. 1–41, 2022.

[5] S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning. Cambridge, MA, USA: MIT Press, 2023.

[6] Pymetrics, "Bias audit and fairness report," 2020. [Online]. Available: https://www.pymetrics.ai [Accessed: Jul. 2, 2025].

[7] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2014, pp. 94–108.

[8] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105–6114.

[9] S. Aslan, U. Güdükbay, and H. Dibeklioğlu, "Multi-modal assessment of apparent personality using feature attention and error consistency constraint," Image Vis. Comput., vol. 110, p. 104163, 2021, doi: 10.1016/j.imavis.2021.104163.

[10] FFmpeg Developers, "FFmpeg," 2021. [Online]. Available: https://ffmpeg.org/ [Accessed: Aug. 2, 2025].

[11] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in Python," in Proc. 14th Python Sci. Conf., 2015, pp. 18–25.

[12] G. Bradski, "The OpenCV library," Dr. Dobb's J. Softw. Tools, vol. 25, no. 11, pp. 120–125, 2000.

[13] H. Kaya, A. A. Salah, and F. Gunes, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," in Proc. CVPR ChaLearn First Impressions Challenge Workshop, 2017.

[14] H. Kaya, F. Gürpınar, and A. A. Salah, "Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video CVs," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2017, pp. 1–9, doi: 10.1109/CVPRW.2017.210.

[15] T. Yang, J. Deng, X. Quan, and Q. Wang, "Orders are unwanted: Dynamic deep graph convolutional network for personality detection," in Proc. AAAI Conf. Artif. Intell., 2023, pp. 13896–13904.

[16] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multi-modal machine learning: A survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.

[17] S. Yan, D. Huang, and M. Soleymani, "Mitigating biases in multi-modal personality assessment," in Proc. Int. Conf. Multi-modal Interaction (ICMI), 2020, pp. 209–217, doi: 10.1145/3382507.3418889.

[18] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2016, pp. 3315–3323.

[19] M. Zafar, I. Valera, M. Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment," in Proc. World Wide Web Conf. (WWW), 2017, pp. 1171–1180, doi: 10.1145/3038912.3052660.

[20] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in Proc. 30th Int. Conf. Mach. Learn. (ICML), vol. 28, no. 3, pp. 325–333, 2013. [Online]. Available: https://proceedings.mlr.press/v28/zemel13.html.

[21] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints," in Proc. 2017 Conf. Empirical Methods Nat. Lang. Process. (EMNLP), Copenhagen, Denmark, Sep. 2017, pp. 2979–2989, doi: 10.18653/v1/D17-1323.

[22] Retorio, "Personality AI for recruiting and coaching," 2021. [Online]. Available: https://www.retorio.com [Accessed: Aug. 2, 2025].

[23] O. Kampman, E. J. Barezi, D. Bertero, and P. Fung, "Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction," in Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (ACL), vol. 2, Melbourne, Australia, Jul. 2018, pp. 606–611. doi: 10.18653/v1/P18-2096.
[24] D. Grimm, "Personality trait prediction," GitHub, 2020. [Online]. Available: https://github.com/grimmdaniel/personality-trait-prediction [Accessed: Aug. 2, 2025].

[25] J.-I. Biel, O. Aran, and D. Gatica-Perez, "You are known by how your vlog: Personality impressions and nonverbal behavior in YouTube," in Proc. Int. Conf. Weblogs Soc. Media (ICWSM), 2011.

[26] J.-I. Biel and D. Gatica-Perez, "The YouTube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs," IEEE Trans. Multimedia, vol. 15, no. 1, pp. 41–55, 2013, doi: 10.1109/TMM.2012.2225032.

[27] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M.

Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, 2016, pp. 3–10, doi: 10.1145/2988257.2988258.

[28] F. Mairesse, M. A. Walker, M. Mehl, and R. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," J. Artif. Intell. Res., vol. 30, pp. 457–500, 2007, doi: 10.1613/jair.2349.

[29] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open-source facial behavior analysis toolkit," in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2016, pp. 1–10, doi: 10.1109/WACV.2016.7477553.

[30] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE – The Munich versatile and fast open-source audio feature extractor," in Proc. 18th ACM Int. Conf. Multimedia, 2010, pp. 1459–1462, doi: 10.1145/1873951.1874246.

[31] R. Bellamy, K. Dey, M. Hind, S. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. Varshney, and Y. Zhang, "AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," IBM J. Res. Dev., vol. 63, no. 4/5, pp. 4:1–4:15, 2019, doi: 10.1147/JRD.2019.2942287.

[32] R. Binns, M. Veale, M. Van Kleek, and N. Shadbolt, “Like trainer, like bot?’ Inheritance of bias in algorithmic content moderation," in Proc. Conf. Fairness, Accountability, Transparency (FAT), 2018, pp. 15–27.

[33] Enzuzo, “OneTrust vs. UpGuard: Key Differences and Comparison [Review],” *Enzuzo*, 2023. [Online]. Available: https://www.enzuzo.com/alternatives/onetrust-vs-upguard. [Accessed: Jul. 3, 2025].