



## **Ensuring Fairness and Responsibility in AI-Based Psychological Assessment: A Review of Bias, Validity, and Ethical Challenges**

Samiksha Kourav

B.Sc. Forensic Science, Department Of Forensic Science, Institute of Sciences

SAGE University, Indore, M.P.

[samkourav1218@gmail.com](mailto:samkourav1218@gmail.com)

### **Abstract**

Artificial intelligence (AI) is increasingly being prevalent in psychological assessment, with machine learning and large language models facilitating for personality assessment, mental health screening, Clinical decision support, analysis of responses, behavioral evaluation, and automated interpretation of psychological data. These advancement's providing valuable benefits, opportunities, including greater efficiency, accessibility, and scalability. However, these increasing advancements by using AI also raises an important question: can an assessment be considered trustworthy, fair, valid, and transparent simply because an AI system produces accuracy ? An assessment may produce appropriate or predictive accuracy while still raising concerns about reliability, fairness, ethical accessibility, and furthermore.

This review examines research published between 2019 and 2026 on Advancements in psychological assessment, focusing on three interconnected areas: fairness and algorithmic bias, psychological validity and reliability, and ethical responsibility. Although existing research has increasingly addressed these concerns, these factors are often frequently examined separately rather than collectively, leaving gaps in existing research in understanding how these factors interact within and affect to psychological assessment. AI systems may demonstrate high overall performance while generating variation in outcomes across diverse backgrounds, cultural, linguistic, or socioeconomic groups. Likewise, systems that provide clear and interpretable decision making processes may still lack sufficient psychometric validity or reliability.

Accordingly, this review evaluates, Relying on data sets that may Risk of over reliance on automated recommendations , and ambiguity in human responsibility. It also considers their potential impacts, including inaccurate or unequal assessment outcomes, Potentially biased psychological judgments and perpetuation of social inequalities.

This review takes a broader approach to look beyond the technical capabilities of advancements to consider its wider implications for psychological assessment while also considering the challenges associated with particular attention to both measurement quality and responsible practice. It considers whether these systems can assess psychological characteristics appropriately, perform fairly across different groups, and produce results that can be meaningfully interpreted. It also highlights the importance of proper validation, representative datasets, transparency, privacy protection, and human involvement throughout the assessment process. Ultimately, the effectiveness of AI-based psychological assessment should be judged from a wider perspective than predictive performance. The assessment process must also demonstrate sound measurement, fairness across individuals and groups, clear interpretation of results, protection of personal information, and appropriate human oversight.

**Keywords:** artificial intelligence, psychological assessment, AI-based assessment, psychometric validity, fairness, algorithmic bias, reliability, transparency, ethical AI, human oversight.

## 1. Introduction

Psychological assessment is as structured and systematic process of collecting, integrating, and interpreting information regarding an individual's behavioural, and personality characteristics, cognitive, emotional. The scientific interpretation of psychological assessment results is fundamentally dependent upon established principles of measurement, particularly reliability, validity, and the appropriateness of an assessment instrument for the population and purpose for which it is used [15]. Recent advances in to advance conventional psychological assessment practices are increasingly being complemented towards data driven approaches. Computational technologies into psychological science, conventional assessment paradigms are progressively demonstrating the ability of analysing linguistic, behavioural, physiological, and multimodal information.

Artificial intelligence (AI), encompassing machine learning (ML), natural language processing (NLP), deep learning, computer vision, and large language models (LLMs), has consequently emerged as a significant methodological development within psychological and mental-health research. AI-based systems are being investigated for applications including mental-health screening, psychopathological symptom detection, personality assessment, behavioural evaluation, automated response analysis, and clinical decision support [16]. A recent scoping review examining AI as a psychological assessment tool identified substantial growth in this field but also reported considerable methodological heterogeneity, including variable sample characteristics, limited demographic and cultural analyses, inconsistent use of established diagnostic standards, and insufficient transparency [16].

The computational capabilities of AI provide several potential advantages for psychological assessment. Automated systems can process large and heterogeneous datasets, identify complex patterns, facilitate rapid analysis, and potentially support scalable and remotely accessible assessment procedures [16,17]. The emergence of LLMs has further expanded these possibilities because their capacity to process natural language permits investigation of psychological constructs through conversational and text-based interactions. Recent psychometric research has demonstrated that certain LLMs can exhibit measurable and, under specific conditions, reliable and valid personality-related characteristics, thereby illustrating the relevance of established psychometric frameworks to AI-based assessment [18]. Nevertheless, technological performance should not be equated with psychometric adequacy.

Psychological assessment involves the measurement of latent constructs that cannot be directly observed and must therefore be inferred from observable indicators. Consequently, an AI system demonstrating high predictive accuracy does not necessarily establish that the system is measuring the intended psychological construct. Evidence concerning construct validity, criterion-related validity, reliability, measurement invariance, generalizability, and clinical utility remains essential before AI-derived outputs can be considered scientifically defensible assessment information [15]. This distinction between computational performance and psychometric validity represents a central issue in the responsible development of AI-assisted psychological assessment.

A further concern is algorithmic bias and fairness. AI systems are dependent upon the composition, quality, and representativeness of their training and validation datasets. When datasets contain demographic imbalance, cultural underrepresentation, linguistic limitations,

or historically embedded social inequities, resulting models may reproduce or amplify existing disparities [19]. In mental-health applications, biased algorithms may potentially contribute to differential classification, inaccurate risk estimation, or unequal access to appropriate assessment and intervention. Therefore, evaluation of AI systems must extend beyond aggregate performance metrics and examine whether model performance and assessment outcomes remain equitable across relevant demographic, cultural, and linguistic groups [19].

The issue of transparency and explainability presents an additional methodological challenge. Complex AI architectures may generate outputs through computational processes that are difficult for users to interpret at the individual-decision level. Limited interpretability can restrict the ability of psychologists to determine why a particular prediction was generated, which features contributed to it, and whether those features possess meaningful psychological significance. Contemporary recent scholarly work underscores or indicates the importance of human oversight such as accountability, validation, transparency as critical considerations in the implementation of AI based approaches into psychological and educational assessment .

The accumulating evidence base further demonstrates a Central research scientific question: How reliably can AI grounded in psychological assessment systems evaluate psychological characteristics and ethically responsible evaluations and what factors influence their validity fairness transparency and ethical acceptability while minimizing bias and preserving human oversight? Although existing research demonstrates considerable technological potential, the evidence base remains heterogeneous, with persistent limitations concerning methodological standardization, external validation, demographic diversity, cultural generalizability, interpretability, and clinically meaningful validation [16,15,17]. Recognizing and mitigating these limitations is necessary before Advanced Computational Intelligence can be responsibly implemented with standard clinical practice, educational, Professional psychological assessment, or forensic framework.

This review systematically Evaluates the evolving landscape of Recent advancing applications of artificial intelligence in Psychometric assessment functioning with particular emphasis on psychometric validity and reliability, algorithmic bias, fairness, explainability, transparency, privacy, accountability, and ethical governance. It Synthesizes the critical appraisal of current research focusing on methodological rigor scientific validity and practical

constraints of AI-assisted assessment and identifies Highlighting research directions including standardized psychometric validation, cross-cultural evaluation, transparent model development, continuous performance monitoring, diverse and representative datasets, and meaningful human oversight. The review further examines and also evaluates for the responsible integration of AI into psychological assessment in a manner that complements established scientific and professional standards rather than compromising them.

## 2. Review Methodology

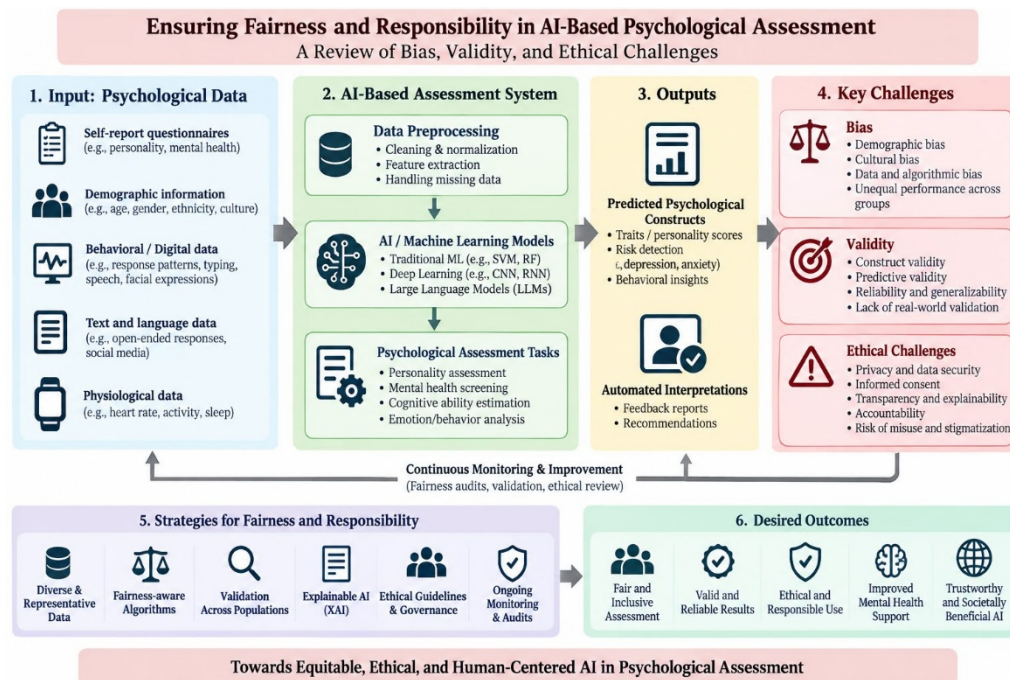


**Fig 1:** Conceptual framework illustrating AI methodologies, psychological assessment applications, evaluation dimensions, ethical challenges, principles for responsible implementation

The present study adopts a qualitative narrative review approach to examine recent developments and challenges associated with AI-based psychological assessment.

### 2.1 Literature Scope

The current research focuses on research published between 2021 and 2026, with emphasis on studies examining AI, machine learning, generative AI, natural language processing, Advanced language models In psychological and mental health screening and evaluation.

The literature was examined across several major areas related to AI-based psychological assessment. These included psychological assessment, mental-health screening

and evaluation, along with machine-learning-based prediction, behavioural analysis, and natural-language analysis. The review also focused on major psychological measurement concerns such as reliability and validity, as well as emerging challenges related to algorithmic bias and fairness. In addition, issues of privacy, transparency, ethical AI, and human AI collaboration were considered to understand how AI can be applied responsibly and effectively in psychological assessment.

Systematic reviews, scoping reviews, empirical studies, methodological research, and authoritative ethical guidance were given particular importance.

## ***2.2 Search Concepts***

The major search concepts included combinations of:

- “artificial intelligence” AND “psychological assessment”
- “machine learning” AND “psychological assessment”
- “AI” AND “psychometric validity”
- “algorithmic bias” AND “mental health”
- “fairness” AND “AI” AND “psychological assessment”
- “digital phenotyping” AND “mental health”
- “large language models” AND “psychological assessment”
- “generative AI” AND “mental health”
- “AI ethics” AND “psychological assessment”

## ***2.3 Analytical Approach***

Rather than quantitatively combining effect sizes, the literature was qualitatively synthesised to identify major developments, methodological limitations, ethical concerns, recurring findings, and Critical avenues for further research.

Research gaps and future directions were identified with emphasis on whether AI systems can measure psychological characteristics appropriately, demonstrate adequate reliability and validity, perform equitably across diverse populations, provide interpretable results, protect sensitive personal information, and maintain appropriate professional and human oversight.

## **3. Development and Applications of AI in Psychological Assessment**

The development of AI-based psychological assessment can be understood as a progression from relatively simple computational classification toward increasingly complex, multimodal, and generative systems.

Early applications largely involved machine-learning models trained to identify relationships between psychological outcomes and measurable characteristics. These models commonly used questionnaires, clinical records, behavioural variables, or linguistic information.

Over time, AI applications expanded to include digital phenotyping, computer vision, multimodal systems smartphone sensing, wearable devices, speech analysis, and natural-language processing, . More recently, generative AI and LLMs have introduced the possibility of assessing psychological information from naturally occurring language and interactive conversations.

The 2026 scoping review by Dev et al. provides evidence of the rapid growth of this field. Of the 320 studies included, 75% used supervised learning and 47% used deep learning, while 77% compared multiple algorithms. The research was concentrated largely on depression, anxiety, suicidality, and stress [6].

This progression demonstrates that AI is increasingly moving from a tool for **automated classification** toward a potential component of the broader psychological measurement process.

However, increased computational sophistication does not necessarily mean increased psychometric quality.

#### **4. Algorithmic Bias and Fairness**

Psychological screening is one of the most common applications of AI.

AI systems have been investigated for identifying individuals who may be experiencing depression, anxiety, stress, suicidal thoughts or behaviours, psychological distress, mood disturbances, and other mental-health-related difficulties.

The principal advantage of AI-based screening is scalability. A computational system may potentially process large numbers of individuals more rapidly than conventional assessment conducted exclusively by professionals.

However, screening is inequivalent to diagnosis.

A screening system identifies an increased probability or risk of a particular condition. It does not necessarily establish a clinical diagnosis. This distinction is particularly important because the 2026 scoping review found that many studies validated AI systems using screening instruments rather than established DSM or ICD diagnostic criteria. Approximately 71% of studies in the review did not use DSM or ICD criteria [6].

Therefore, AI-based psychological screening should generally be interpreted as an assessment-support mechanism, rather than as an autonomous diagnostic process.

## **5. Machine Learning and Psychological Prediction**

Machine learning allows computational systems to identify statistical relationships between predictor variables and psychological outcomes.

Commonly used approaches include:

- › logistic regression;
- › support vector machines;
- › random forests;
- › gradient-boosting methods;
- › artificial neural networks;
- › deep-learning models; and
- › ensemble models.

These approaches have been applied to psychological screening and prediction using clinical, behavioural, demographic, linguistic, and digital information.

The advantage of machine learning is its capacity to analyse complex interactions between variables that may be difficult to identify using conventional statistical approaches.

However, the same complexity can create difficulties in interpretation.

A model may identify a statistically useful relationship without providing a psychologically meaningful explanation for that relationship. Consequently, machine-learning prediction must be distinguished from psychological understanding.

## **6. Digital Phenotyping and Continuous Assessment**

Digital phenotyping represents a major development in AI-assisted psychological assessment.

It involves the collection and analysis of data generated through everyday interaction with digital devices. Such data may include:

- › smartphone activity;
- › sleep;
- › physical activity;
- › mobility;
- › location patterns;
- › screen interaction;
- › communication behaviour;
- › wearable-sensor information; and
- › physiological indicators.

The principal attraction of digital phenotyping is that it can potentially provide repeated or continuous information rather than relying on a single assessment session.

This may improve the ecological relevance of assessment because psychological functioning can be examined in everyday environments.

However, digital behaviour should not automatically be interpreted as a direct indicator of psychological functioning.

For example, reduced smartphone activity may be associated with depressive symptoms, but it could also result from employment patterns, academic workload, technological limitations, social circumstances, or personal preference.

Research published in 2024 demonstrated that sensed behavioural predictors of depression varied across demographic and socioeconomic subgroups, suggesting that the same digital signal may not have the same psychological meaning across individuals [8].

Thus, digital phenotyping should be considered a contextual source of psychological information, rather than an inherently objective measurement of mental health.

## **7. Algorithmic Bias in AI-Based Psychological Assessment**

Algorithmic bias is one of the most important challenges in AI-based assessment.

Bias may originate from several stages of the assessment process:

### ***7.1 Sampling Bias***

The training sample may not adequately represent the population in which the system will eventually be used.

### ***7.2 Representation Bias***

Certain demographic or cultural factors may be marginally represented in the dataset.

### ***7.3 Measurement Bias***

The variables used by an AI model may not represent psychological constructs equally across groups.

### ***7.4 Label Bias***

The outcome used to train the model may itself contain errors or systematic differences between populations.

### ***7.5 Algorithmic Bias***

The model may generate systematically different outcomes for different groups.

### ***7.6 Deployment Bias***

A model developed for one population or setting may be used in another population without appropriate validation.

These forms of bias demonstrate that AI does not automatically eliminate human or social bias. Instead, AI can reproduce and could exacerbate existing biases present in the data and systems from which it learns.

## **8. Evidence of Bias Across Demographic Groups**

Recent empirical research provides direct evidence that algorithmic fairness must be evaluated rather than assumed.

Dang et al. [\[7\]](#) investigated machine-learning models for depression prediction across four populations and found biased behaviour associated with demographic and socioeconomic characteristics. Their results indicated that Established machine learning techniques could produce inequitable outcomes across protected groups. The Previous researchers further demonstrated that Algorithmic bias reduction approach could reduce some forms of discrimination, although no single model provided universally equal outcomes.

These findings have important Consequences for psychological assessment.

A model may demonstrate good overall accuracy while simultaneously producing higher false-positive or false-negative rates for particular groups.

Therefore:

Overall model performance cannot be considered sufficient evidence of fairness.

Fairness assessment should include subgroup-specific examination of performance.

Relevant indicators may include:

- sensitivity;
- specificity;
- false-positive rate;
- false-negative rate;
- positive predictive value;
- negative predictive value;
- calibration; and
- equality of opportunity or related fairness measures.

A 2026 scoping review of fairness metrics in clinical predictive AI found considerable fragmentation in how fairness is defined and measured, with limited clinical validation and substantial reliance on threshold-dependent measures [\[11\]](#).

## **9. Cultural and Linguistic Bias**

Psychological assessment is moderated by linguistic cultural variations.

The expression of social behaviour, emotional distress, personality characteristics, and psychological symptoms may differ across cultural and linguistic contexts.

An AI system trained primarily using English-language or Western populations may therefore not demonstrate equivalent validity when applied to other populations.

This concern is Specifically relevant in diverse countries such as India.

For example, an NLP system may interpret:

- indirect emotional expression;
- culturally specific idioms;
- code-switching;

- regional expressions;
- culturally appropriate communication styles; or
- differences in emotional vocabulary

differently from the way a human professional familiar with the cultural context would interpret them.

The 2026 scoping review identified insufficient demographic and cultural analyses as a continuing weakness in the existing AI psychological-assessment literature [6].

Therefore, cultural fairness should be considered during:

*dataset construction → model development → validation → interpretation → deployment.*

## **10. Digital Inequality and Socioeconomic Bias**

AI-based assessment can also reproduce socioeconomic inequalities.

Digital phenotyping often depends on access to:

- smartphones;
- wearable devices;
- stable internet;
- digital literacy;
- sufficient technological infrastructure.

Individuals while avoiding access to these resources may be disproportionately represented in existing AI research or may generate incomplete datasets.

Furthermore, behavioural patterns can differ according to employment, income, education, housing conditions, and access to healthcare.

Consequently, an AI model may interpret socioeconomic differences as psychological differences. This creates a serious concern because an apparently objective algorithm may unintentionally reproduce existing social inequalities.

The WHO has similarly warned that AI systems developed primarily using data from high-income settings may not perform appropriately in low- and middle-income settings, emphasizing the need for diversity and equity in AI development [9].

## **11. Psychological Validity of AI-Based Assessment**

Validity is central to psychological assessment.

An assessment must demonstrate that it measures what it is intended to measure and that the interpretation of its results is supported by empirical evidence.

AI-based psychological assessment should therefore be evaluated using established forms of validity, including construct validity, criterion-related validity, convergent validity, discriminant validity, ecological validity, incremental validity, and measurement invariance. High predictive accuracy alone does not establish these forms of validity, as an AI system may achieve strong prediction while still failing to accurately measure the intended psychological construct or perform consistently across different populations and contexts.

## **12. Construct Validity**

Construct validity is particularly important because AI may identify a statistical correlate of a psychological characteristic without actually measuring the psychological construct.

For example, an AI system may predict depression from reduced physical activity.

However, reduced physical activity can also occur because of:

- physical illness;
- unemployment;
- environmental restrictions;
- injury;
- lifestyle preferences; or
- socioeconomic circumstances.

Thus, the model may be able to estimate and identify indicators associated with depression outcomes without necessarily measuring depression itself.

This distinction should remain central when interpreting AI-driven psychological assessment.

## **13. Criterion and Diagnostic Validity**

Criterion validity requires comparison with an appropriate external criterion.

In psychological assessment, this may include:

- structured clinical interviews;
- clinician-rated assessment;

- › established psychometric instruments;
- › DSM criteria;
- › ICD criteria; and
- › clinically meaningful longitudinal outcomes.

The 2026 review by Dev et al. identified limited use of DSM and ICD diagnostic criteria across the literature, with many studies relying on screening instruments or algorithmic cross-validation [6].

This raises a significant concern:

If an AI system is trained to predict a screening questionnaire, successful prediction of that questionnaire does not necessarily demonstrate successful diagnosis.

Future research should therefore distinguish between:

*questionnaire prediction* → *symptom classification* → *screening* → *clinical diagnosis*.

These are not interchangeable outcomes.

## 14. Reliability of AI-Based Psychological Assessment

Reliability refers to the consistency and stability of measurement.

For AI-based psychological assessment, reliability may involve test retest stability, temporal consistency, consistency across devices model stability, inter-method agreement, and demographic groups, and reproducibility of AI-generated outputs. These aspects are important for ensuring that AI-based assessments produce consistent and dependable results across different time points, settings, technologies, and populations.

Generative AI creates additional reliability concerns because its outputs may vary depending on prompt formulation, model version, input context, system configuration, temperature or generation settings, and changes made by model providers. Such variability can affect the consistency and reproducibility of AI-based psychological assessments, making it important to evaluate whether similar inputs produce stable results under comparable conditions.

Therefore, an AI system cannot be considered reliable merely because the underlying computational procedure is automated.

Reliability must be empirically demonstrated.

## 15. External Validation and Generalisability

One of the major weaknesses identified in AI research is limited external validation.

A model may perform well when tested on data similar to those used during development but perform poorly in a new population.

The 2023 systematic review by Tornero-Costa et al. identified substantial methodological weaknesses in AI mental-health research, including inadequate preprocessing reporting, limited external validation, poor transparency, and insufficient reproducibility [12].

This creates an important distinction:

Internal validation  $\neq$  External validity

Cross-validation can demonstrate how a model performs within a particular dataset, but it cannot fully establish whether the model will perform appropriately in another hospital, university, geographical region, language, culture, or demographic population. Therefore, external validation across diverse settings and populations is essential to determine the generalizability, fairness, and reliability of AI-based psychological assessment systems.

External validation should therefore become an essential component of responsible AI assessment research.

## 16. Predictive Accuracy Versus Assessment Quality

Predictive accuracy is often presented as the principal evidence of AI performance.

Common measures include:

- accuracy;
- sensitivity;
- specificity;
- precision;
- F1 score;
- AUC;
- mean absolute error; and
- calibration.

These measures are useful but incomplete.

A model may achieve a high AUC while:

- › relying on biased variables;
- › performing poorly for minority groups;
- › lacking construct validity;
- › being poorly calibrated;
- › lacking external validation; or
- › being clinically difficult to interpret.

The 2026 scoping review found AUC to be the most frequently reported performance metric in AI psychological-assessment studies [6].

This suggests that future studies should adopt a broader evaluation framework.

A more appropriate model of assessment quality is:

*Assessment Quality = Accuracy + Reliability + Validity + Fairness + Interpretability + Clinical Utility + Ethical Responsibility*

## 17. Large Language Models and Psychological Assessment

Advanced Large language models represent one of the most significant developments in contemporary AI.

LLMs can process extensive amounts of natural language and identify linguistic patterns associated with human behaviour and psychological characteristics.

Kjell et al. [10] argued that LLMs may potentially expand psychological assessment beyond conventional rating scales by analysing naturally occurring language, provided that their use is subjected to appropriate empirical evaluation.

This creates a potential shift from:

*Structured questions → fixed responses → numerical scores*

toward:

*Natural communication → computational analysis → psychological inference*

The latter approach may provide richer information about how people naturally communicate.

However, it also introduces additional uncertainty.

Natural language is contextual, culturally influenced, interpersonal relationships, ambiguous, dynamic, dependent on and influenced by education and linguistic ability. Therefore, linguistic fluency in AI-generated or AI-analysed language does not by standalone establish psychological validity, as accurate interpretation requires consideration of the individual's cultural, social, linguistic, and contextual background.

## **18. Generative AI and Contemporary Ethical Challenges**

Generative AI has expanded rapidly in mental-health-related applications.

A 2025 systematic review examined 783 records and included 79 studies concerning generative AI in mental health. Nearly half of the included studies addressed diagnosis or assessment, particularly detection of depression and suicidality through text data [\[13\]](#).

The same review identified recurring concerns involving:

- › privacy and security;
- › information integrity;
- › fairness;
- › user safety;
- › ethical governance; and
- › responsible engagement [\[13\]](#).

These findings demonstrate that generative AI introduces both opportunities and risks.

AI may assist with the interpretation of large quantities of psychological information, but generated outputs can also appear authoritative even when they are incorrect or insufficiently supported.

Therefore:

Human-like communication should not be confused with clinically valid psychological judgement.

## **19. Privacy and Confidentiality**

Psychological assessment involves highly sensitive information.

AI-based psychological assessment may involve a wide range of sensitive data, including personal narratives, clinical records, speech, facial information, behavioural patterns, location data, mental-health symptoms, smartphone use, physiological information, and social-media activity. The combination of these data sources may allow the creation of highly detailed psychological profiles, increasing the potential risks associated with privacy and data misuse. Consequently, privacy protection must be considered throughout the entire AI assessment lifecycle. The principles include informed consent, data minimisation, secure storage, restricted access, appropriate anonymisation, transparent data-use policies, clear data-retention periods, and protection against unauthorised secondary use.

The WHO's 2021 guidance specifically identifies privacy, confidentiality, informed consent, human autonomy, transparency, and accountability as core considerations in responsible AI for health [\[9\]](#).

## **20. Informed Consent in AI-Based Assessment**

Traditional psychological assessment generally involves informing individuals about the purpose of assessment and how their information will be used.

AI introduces additional questions:

- › Is the person aware that AI is involved?
- › What information is being collected?
- › Is the information used to train future models?
- › Who has access to the data?
- › Can the person challenge an AI-generated result?
- › How long will the data be stored?
- › Can the individual withdraw consent?

Meaningful consent therefore requires more than simply informing an individual that “AI is being used.”

Individuals should understand the role, limitations, potential risks, and consequences of AI-based assessment.

## **21. Transparency and Explainability**

Transparency is another major requirement for responsible AI-based psychological assessment.

A psychologist or individual receiving an AI-generated assessment result should ideally be able to understand what the system was designed to assess, what data were used, how the model was validated, what limitations exist, how confident the prediction is, and whether the model has been evaluated for the relevant population. Such transparency is essential for interpreting AI-generated results appropriately, identifying potential limitations or biases, and supporting informed professional decision-making.

The WHO's 2021 framework identifies transparency, explainability, and intelligibility as central principles for AI in health [\[9\]](#).

However, explainability standalone does not establish validity.

An AI system can provide a convincing explanation for a prediction while still relying on biased information or inappropriate.

Therefore:

Explainability must accompany instead of replacement of psychometric validation.

## **22. Human Oversight and Professional Responsibility**

One of the most important ethical principles in AI-based psychological assessment is the preservation of human responsibility.

AI can assist with:

- › data processing;
- › pattern recognition;
- › automated scoring;
- › screening;
- › information organisation;
- › longitudinal monitoring; and
- › preliminary interpretation.

However, qualified professionals remain essential in AI-based psychological assessment because they provide contextual interpretation, clinical judgement, ethical decision-making, effective communication, consideration of individual circumstances, differential

interpretation, and final professional responsibility. AI systems can support assessment processes, but professional oversight is necessary to ensure that results are interpreted appropriately and that decisions remain accurate, ethical, and sensitive to the needs of the individual.

The WHO emphasizes that humans should remain in control of health-related decision-making and that appropriate accountability mechanisms should exist when algorithmic systems cause harm [9].

Accordingly, AI should generally be viewed as an augmentative system rather than an autonomous psychological assessor.

## 23. Automation Bias and Over-Reliance on AI

The increasing use of AI also creates the possibility of automation bias.

Automation bias occurs when professionals place excessive trust in computer-generated recommendations and fail to adequately consider contradictory evidence.

In psychological assessment, this could occur when:

*AI prediction → professional acceptance → psychological judgement*

replaces:

*AI prediction → critical evaluation → contextual interpretation → professional judgement*

This distinction is crucial.

An AI-generated risk score may appear objective because it is expressed numerically. However, numerical precision does not necessarily represent psychological certainty.

Therefore, professionals should be trained to critically evaluate AI outputs rather than automatically accept them.

## 24. Accountability and Legal Responsibility

AI-based assessment raises an important question:

Who is responsible when an AI assessment is wrong?

Potentially relevant stakeholders in AI-based psychological assessment include AI developers, technology companies, psychologists, clinicians, institutions, researchers, data

scientists, and regulatory authorities. Each stakeholder has an important role in ensuring that AI-based assessment systems are developed, evaluated, implemented, and regulated in a manner that promotes accuracy, fairness, transparency, privacy, and ethical responsibility.

Responsibility should not become ambiguous simply because an algorithm contributed to a decision.

The WHO recommends clear accountability mechanisms and the ability to question and seek redress for harmful algorithmic decisions [9].

For psychological assessment, this implies that professional responsibility must remain clearly identifiable.

## 25. Fairness as a Psychometric Issue

Fairness should not be considered only as an ethical issue.

It is also a measurement issue.

If an assessment does not measure a psychological construct equivalently across groups, its validity is compromised.

For example, if an AI model accurately predicts depression in one demographic group but systematically underestimates depression in another group, the model cannot be considered equally valid across those populations.

Therefore:

*Fairness → Measurement equivalence → Validity → Responsible interpretation*

These concepts are interconnected rather than independent.

This represents an important conceptual gap in existing research, where fairness, validity, and ethics are frequently investigated separately.

## 26. Interaction Between Bias, Validity and Ethics

The central argument of this review is that bias, validity, and ethical responsibility should not be examined independently.

These dimensions interact throughout the assessment process.

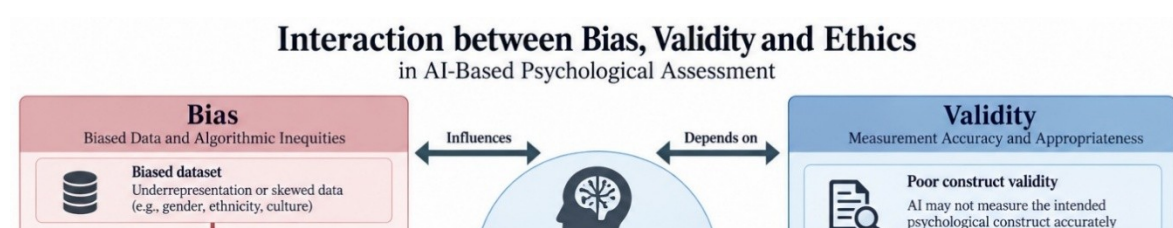


**Fig 2:** Interaction between Bias, Validity, and Ethics in AI-Based Psychological Assessment.

## 27. Current Evidence From 2021–2026

The literature reviewed during the period demonstrates a clear development in research priorities.

| Period | Major research development                           | Major concern                                                     |
|--------|------------------------------------------------------|-------------------------------------------------------------------|
| 2021   | AI governance and machine-learning applications      | Human autonomy, privacy, safety and equity                        |
| 2022   | Expansion of digital and computational assessment    | Measurement quality and generalisability                          |
| 2023   | Digital phenotyping and methodological evaluation    | External validation and transparency                              |
| 2024   | Greater focus on algorithmic fairness and LLMs       | Demographic bias and cultural validity                            |
| 2025   | Rapid expansion of generative AI                     | Privacy, fairness, safety and information integrity               |
| 2026   | Large-scale synthesis of AI psychological assessment | Limited diagnostic validation, cultural analysis and transparency |

The most recent 2026 scoping review is particularly significant because it mapped 320 empirical studies and demonstrated both the rapid growth and methodological limitations of the field [\[6\]](#).

## 28. Research Gaps

Despite rapid development, important gaps remain.

### ***28.1 Limited Population Diversity***

AI research remains concentrated in particular geographical regions and populations. The 2026 review reported substantial representation from China and the United States, indicating a need for greater international diversity [6].

### ***28.2 Insufficient Cultural Validation***

Many systems have not been adequately tested across cultures and languages.

### ***28.3 Limited External Validation***

Numerous models rely on internal cross-validation rather than independent external samples.

### ***28.4 Reliance on Self-Report***

AI systems frequently use self-report questionnaires as outcome measures, potentially limiting conclusions about clinical validity [6].

### ***28.5 Limited Use of Established Diagnostic Standards***

The limited use of DSM and ICD criteria creates uncertainty regarding clinical diagnostic validity [6].

### ***28.6 Lack of Standardised Fairness Assessment***

Research uses different definitions and metrics of fairness, making comparisons difficult [11].

### ***28.7 Limited Longitudinal Research***

More research is required to establish whether AI can reliably monitor psychological changes over time.

### ***28.8 Limited Reproducibility***

Rapid changes in AI models, particularly generative AI, create challenges for reproducibility.

## **29. Recommendations for Responsible AI-Based Psychological Assessment**

Based on the literature reviewed, several recommendations can be proposed.

### ***29.1 Representative Datasets***

AI systems should be trained and validated using diverse populations representing the intended users.

### ***29.2 Psychometric Validation***

AI systems should be evaluated using:

- reliability;
- construct validity;
- criterion validity;
- convergent validity;
- discriminant validity;
- measurement invariance; and
- clinical utility.

### ***29.3 Subgroup Performance Analysis***

Researchers should report performance across relevant demographic and cultural groups rather than reporting only overall accuracy.

### ***29.4 External Validation***

Models should be tested independently across different populations and settings.

### ***29.5 Cultural and Linguistic Validation***

Assessment systems should undergo culturally and linguistically appropriate validation rather than relying solely on translation.

### ***29.6 Transparent Reporting***

Researchers should clearly report:

- data sources;
- sample characteristics;
- model development;
- validation procedures;
- performance metrics;
- subgroup performance;
- limitations; and
- intended use.

### ***29.7 Human Oversight***

High-stakes psychological decisions should remain under appropriate professional supervision.

### ***29.8 Privacy Protection***

Data collection should follow principles of minimisation, confidentiality, security, informed consent, and appropriate access control.

### ***29.9 Continuous Monitoring***

AI systems should be periodically evaluated after deployment to identify performance changes and emerging biases.

## **30. Proposed Framework for Fair and Responsible AI-Based Psychological Assessment**

The findings of this review support a Fairness–Validity–Responsibility Framework.

### ***Stage 1: Define the Psychological Construct***

Clearly identify what the AI system is intended to assess.

↓

### ***Stage 2: Develop Representative Data***

Ensure adequate demographic, cultural, linguistic, and socioeconomic representation.

↓

### ***Stage 3: Assess Data Quality and Bias***

Identify sampling, measurement, representation, and label bias.

↓

### ***Stage 4: Develop the AI Model***

Use appropriate computational methods based on the psychological question.

↓

### ***Stage 5: Psychometric Validation***

Evaluate reliability, construct validity, criterion validity, and measurement invariance.

↓

***Stage 6: Fairness Evaluation***

Examine performance across demographic, cultural, linguistic, and socioeconomic groups.

↓

***Stage 7: External Validation***

Test the model using independent populations and real-world settings.

↓

***Stage 8: Transparency and Explainability***

Communicate model performance, limitations, uncertainty, and intended use.

↓

***Stage 9: Human Professional Oversight***

Ensure qualified professionals interpret AI outputs and retain responsibility for high-stakes decisions.

↓

***Stage 10: Ethical Governance and Continuous Monitoring***

Protect privacy, confidentiality, autonomy, fairness, and accountability throughout the AI lifecycle.

This framework emphasises that fairness and ethical responsibility should be incorporated from the beginning of AI development rather than added after the system has been completed.

## **31. Future Directions**

Future research should move from proof-of-concept AI models toward rigorous, psychologically grounded assessment systems.

First, research should prioritise diverse and representative datasets. AI systems intended for global or multicultural use should not rely predominantly on data from a limited number of countries or populations.

Second, psychological researchers should increasingly incorporate psychometric principles into AI development. Predictive accuracy should be considered one component of assessment quality rather than the final criterion.

Third, future systems should combine multiple forms of information. A potentially useful model may integrate:

*validated psychological tests + clinical interview + behavioural information + digital phenotyping + language analysis + longitudinal monitoring + professional judgement.*

Fourth, research should focus on human-AI collaboration rather than complete automation. AI can assist professionals by processing information and identifying patterns, while psychologists retain responsibility for interpretation and ethical decision-making.

Fifth, generative AI requires specific research into:

- › reproducibility;
- › hallucination;
- › prompt sensitivity;
- › cultural bias;
- › privacy;
- › psychological validity;
- › model changes; and
- › human overreliance.

WHO's guidance on large multimodal models further emphasizes that generative AI requires specific governance because such systems can produce outputs that appear human-like and authoritative while introducing risks concerning bias, misinformation, privacy, safety, and accountability [\[14\]](#).

## **32. Discussion**

The literature from 2021 to 2026 demonstrates that AI has considerable potential to transform psychological assessment. Machine learning, digital phenotyping, NLP, multimodal sensing, and generative AI can expand the amount and type of information available for psychological evaluation.

However, the evidence also demonstrates that AI-based assessment cannot be evaluated solely through technical performance.

The 2026 scoping review by Dev et al. provides strong evidence for this conclusion. Despite substantial growth in the number of AI psychological-assessment studies, the field remains characterised by heavy reliance on self-report measures, limited use of established diagnostic standards, insufficient demographic and cultural analysis, and limited transparency [6].

This suggests that technological development has progressed faster than the establishment of comprehensive assessment standards.

Algorithmic fairness provides another important example. Dang et al. [7] demonstrated that machine-learning models predicting depression can produce unequal outcomes across populations. Their findings show that even technically successful AI models may produce biased results unless fairness is explicitly assessed and addressed.

Similarly, smartphone-based depression prediction research demonstrated that behavioural indicators were not consistently associated with depression across demographic and socioeconomic groups [8]. This is particularly important because it challenges the assumption that digital behavioural data are inherently objective.

The findings collectively suggest that fairness is closely connected to validity. If an AI system interprets the same behaviour differently across groups because of differences in training data, its measurement properties may not be equivalent across those groups.

The emergence of generative AI introduces further complexity. LLMs can analyse natural language and may provide new approaches to psychological assessment [10]. However, their ability to generate fluent and apparently authoritative responses can create the risk that users or professionals overestimate their psychological competence.

Therefore, the future of AI-based psychological assessment should not be defined as replacing human psychological expertise.

Instead, a human-centred model of AI-assisted assessment is more appropriate.

In such a model:

AI provides computational capability

while

psychologists provide psychological theory, contextual interpretation, ethical judgement, and professional accountability.

The objective should therefore not be to create an autonomous machine that replaces psychological assessment, but to develop systems that enhance the ability of qualified professionals to conduct scientifically sound, fair, efficient, and responsible assessments.

### **33. Conclusion**

Artificial intelligence is increasingly transforming psychological assessment through machine learning, natural language processing, digital phenotyping, multimodal sensing, and generative AI. Research published between 2021 and 2026 demonstrates considerable potential for improving the efficiency, scalability, accessibility, and continuity of psychological assessment.

However, the evidence also demonstrates that accuracy alone cannot establish that an AI-based assessment is trustworthy.

AI systems may demonstrate high overall predictive performance while producing unequal outcomes across demographic, cultural, linguistic, or socioeconomic groups. Similarly, an AI system may generate reliable predictions without demonstrating adequate construct validity, clinical validity, or generalisability. These concerns are particularly important because psychological assessment involves decisions that may directly influence diagnosis, treatment, education, employment, forensic evaluation, and access to psychological services.

The literature further demonstrates that ethical responsibility cannot be separated from measurement quality. Privacy, confidentiality, informed consent, transparency, accountability, human autonomy, and fairness must be incorporated throughout the development and implementation of AI-based assessment systems.

The most appropriate approach is therefore not to reject AI-based psychological assessment, but to develop and implement it responsibly.

Future systems should be based on representative datasets, robust psychometric validation, subgroup fairness analysis, external validation, cultural responsiveness, transparent reporting, privacy protection, and meaningful human oversight.

Ultimately, AI-based psychological assessment should be evaluated according to a broader standard than predictive performance. A responsible assessment must demonstrate:

accuracy, reliability, validity, fairness, interpretability, privacy, clinical usefulness, and ethical accountability.

The future of psychological assessment is therefore unlikely to depend on a choice between human professionals and artificial intelligence. Rather, it will depend on the development of responsible human-AI collaboration, in which technological innovation strengthens psychological assessment while preserving the scientific principles, ethical standards, professional judgement, and human responsibility that define trustworthy psychological practice.

## References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). [\*Standards for educational and psychological testing\*](#).
- American Psychological Association. (2026). [\*Responsible use of AI in assessment: A practical overview\*](#). American Psychological Association.
- Dang, V. N., Cascarano, A., Mulder, R. H., Cecil, C., Zuluaga, M. A., Hernández-González, J., et al. (2024). Fairness and bias correction in machine learning for depression prediction across four study populations. [\*Scientific Reports, 14, 7848\*](#).
- Dev, V., Consedine, N. S., Gao, Y., Narayanasamy, R., & Serlachius, A. (2026). A scoping review of the use of artificial intelligence as a psychological assessment tool. [\*Translational Psychiatry, 16, 445\*](#).
- Dev, V., Consedine, N. S., Gao, Y., Narayanasamy, R., & Serlachius, A. (2026). [\*A scoping review of the use of artificial intelligence as a psychological assessment tool\*](#). *Translational Psychiatry*, 16, 445.
- Galatzer-Levy, I. R., & Onnela, J.-P. (2023). Machine learning and the digital measurement of psychological health. [\*Annual Review of Clinical Psychology, 19, 133–154\*](#).
- Kjell, O. N. E., Kjell, K., & Schwartz, H. A. (2024). Beyond rating scales: With targeted evaluation, large language models are poised for psychological assessment. [\*Psychiatry Research, 333, 115667\*](#).
- Matos, J., et al. (2026). Critical appraisal of fairness metrics for artificial intelligence-based clinical prediction models: A scoping review. [\*The Lancet Digital Health\*](#).
- Measuring algorithmic bias to analyze the reliability of AI tools that predict depression risk using smartphone sensed-behavioral data.* (2024). [\*npj Mental Health Research\*](#).
- Moura, I., Teles, A., Viana, D., Marques, J., Coutinho, L., & Silva, F. (2023). Digital phenotyping of mental health using multimodal sensing of multiple situations of interest: A systematic literature review. [\*Journal of Biomedical Informatics, 138, 104278\*](#).
- Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. (2025). A psychometric framework for evaluating and shaping

personality traits in large language models. [\*Nature Machine Intelligence\*, 7, 1954–1968.](#)

- Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., Comer, J. S., Brewer, L. C., Frazier, S. L., Chaspari, T., et al. (2023). A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. [\*Perspectives on Psychological Science\*, 18\(5\), 1062–1096.](#)
- Tornero-Costa, R., Martinez-Millana, A., Azzopardi-Muscat, N., Lazeri, L., Traver, D., & Novillo-Ortiz, D. (2023). Methodological and quality flaws in the use of artificial intelligence in mental health research: Systematic review. [\*JMIR Mental Health\*.](#)
- Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P., Carvalho, A. F., Keshavan, M., Linardon, J., & Firth, J. (2021). The growing field of digital psychiatry: Current evidence and the future of mental health care. [\*World Psychiatry\*, 20\(3\), 318–335.](#)
- Torous, J., Wisniewski, H., Bird, B., Carpenter, E., Loeh, K., Martinchek, M., et al. (2023). Digital phenotyping for monitoring mental disorders: Systematic review. [\*Journal of Medical Internet Research\*.](#)
- Wang, X., Zhou, Y., & Zhou, G. (2025). The application and ethical implication of generative AI in mental health: Systematic review. [\*JMIR Mental Health\*.](#)
- World Health Organization. (2021). [\*Ethics and governance of artificial intelligence for health: WHO guidance.\*](#)
- World Health Organization. (2025). [\*Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models.\*](#)