
Ethical Risks of AI-Based Personality Assessment Tools

Dr. Amelia C. Hartwell

Associate Professor, Department of Behavioral Analytics and Responsible Artificial Intelligence, Western Lakes University of Social and Computational Sciences, Minneapolis, Minnesota, USA

Corresponding Author: Dr. Amelia C. Hartwell

Associate Professor, Department of Behavioral Analytics and Responsible Artificial Intelligence, Western Lakes University of Social and Computational Sciences, Minneapolis, Minnesota, USA

Article Type: Simulation-Based Research Article

Submitted: 10-April-2025 **Revised:** 17- April -2025

Accepted: 20- April -2025 **Published:** 24- April -2025

Abstract: Artificial intelligence has grown the range of possible places to infer personality from text, social-media activity, digital footprints, images, media, and other behavioral signals. Existing computational work showed that digital data and social-media language can hold signals for well validated psychological constructs, the Big Five. These new tools have fueled a surge of interest in AI-based personality assessment for employment, marketing, consumer modeling, tailored media, education, and behavioral science. But automated inference amplifies ethical challenges when predictions are used as objective psychological measurements for charitable purposes or important decisions amid questionable construct validity, context dependence, demographic or cultural bias, privacy invasion, weak consent, and low mental capacity for rebuttal.

This concern is heightened for large jobs of high-stakes decisions, in which inferred personality trackers determine who gets hired, fired, ranked, or granted WiFi access. This article describes a new simulation-based Responsible-AI framework for personality assessment we constructed with construct validation, predictive validity, reference-assessment reliability, intergroup error audit, data optimization, pervasive consent, interpretability, purpose limitation, human oversight, and contestability. We sent a synthetic and highly-controlled 2,400-episode audit (a simulated job of personality assessment) equally across four architectures: Standardized Self-Report, Inference-First AI, Explainable AI, and Responsible AI. We used zero real employee, job applicant, student, consumer, patient, social-media, psychological, or commercial assessment data.

The Responsible-AI design achieved an inferred reference-agreement of 0.86, subgroup error disparity of 2.8 points, and high-stakes assessment error of 3.6%. Inference-First AI held similar predictive accuracy but had substantially greater nonconsensual inference, subgroup error disparity, and high-stakes error. Out-of-context signals, weak consent, unstable multimodal sources, and automatic high-stakes use both increased and decreased modelled ethical risk, while construct validation, data minimization, explanation, and limited human review protected it. Results affirm that accuracy alone cannot prove ethical soundness for personality prediction. Instead, person fits, responsible scope of inference, fully negotiated consent, neutral proxies, and a line in the sand separating real risk decisions from inferential shot-calling are all necessary elements of responsible personality prediction.

Keywords: artificial intelligence; personality evaluation; Big Five; personality computing; algorithmic bias; psychological profiling; online traces; privacy; responsible AI; automated recruitment

Introduction

a) Background

The evaluation of personality has become a central concern of a variety of areas within psychology, including the experimental study of social/behavioral traits, personality assessment, individual selection in business and the workplace, and more recently counseling. Among the most widely accepted systems of traits are the so-called Big Five or five-factor models. This five-factors approach is built around several broad areas of personality - commonly referred to as openness to experience, conscientiousness, consciousness, extraversion, agreeableness, and neuroticism or emotional stability. Goldberg proposed the first successful five-factor model, while later work in organizations attempted to relate specific traits to job-performance criteria.

The emergence of Big Data on digital behavior caused a momentous methodological turn. Rather than relying on individual responses to structured instruments, psychologists started inferring a person's personality from their autonomous digital footprints. Kosinski, Stillwell, and Graepel provided the first evidence that inferences of private attributes and

characteristics could be drawn from Facebook ‘Likes’, revealing that even the most mundane behavioral data has the potential to reveal far more personal insights than the individual has directly revealed. Park and colleagues later established that language on social-media can be used as an indicator for the Big Five, comparing them to self-report and informant assessment of traits.

In addition to text, statistical analysis was applied on Twitter, Reddit, photographs and multimodal signals. Plank and Hovy used Twitter for the personal inference, and Gjurković and Šnajder for Reddit. Segalin et al. showed how social context, such as Facebook profile images, could be used for personal inference within personality computing.

The above developments demonstrate technological feasibility. But feasibility and psychological validity are two different things. An AI system might be able to find statistical correlations between behavioral traces and the prescriptive label, but doesn’t guarantee that the resulting prediction is stable enough, equal, transparent or suitable for important personal decisions.

b) Problem Statement

The ultimate moral dilemma of the problem is that AI-powered personality testing is able to make probabilistic links of behavior look like definitive label of the individual. When you might get the classification “high conscientiousness” or “low extraversion”, they would look psychologically specific even if the prediction were highly based on the platform, the language, the recording conditions, the training participants or the behavioral circumstance where it were inferred from.

When this issue arises in the context of using personality assessment to inform employment decisions, the stakes are particularly high. Personality has played a prominent role in personnel psychology for decades, and research has found relationships between certain traits and some job-performance criteria. Yet these studies provide no evidence that every AI-produced personality report from video, face movement, text, or pixels would have the same validity. A landmark meta-analysis by Barrick and Mount found that some relationships deviated from the overall positive effect depending on the given personality dimension, occupation, and criteria.

Algorithmic employment assessment raises more validation and bias issues. Raghavan and colleagues found a great deal of variation in hiring automation vendors' description of the methods used by the algorithm they studied for data collection, validation, and bias mitigation and highlighted the need for the assessment of algorithmic hiring practices to be read from a technical as well as legal perspective. A study published in 2020 also specifically explored bias mitigation in multimodal apparent-personality and hireability assessment as a specific case and showed how demographic bias was already on the table of an algorithmic system assessing applicant personality.

Privacy creates another significant issue. Information that was never intended for psychoanalysis can still reveal an individual's personality. Consequently, behavior in the digital realm can result in a second, unintended form of **psychological** deduction, as a person may consent to uploading images, sending messages, or using a computer program without advertisers' having obtained true informed consent for psychological evaluation.

The major ethical risks addressed in this study include:

- (weak construct validity
- Incorrect personality judgments
- Mixing up the appearance of a personality with the actual personality
- Based on predicting a sample from the population, demographic or cultural prediction gap
- Gaining an overwhelming amount of behavioral information
- Personal assessment without any real consent
- Uses digital footprints for a process of secondary usage
- opaque scoring
- Overconfident interpretation of probabilistic characteristics
- Harassment in the workplace or education programs
- Can't change an inaccurate profile
- Automated high-stakes decisions derived from personality inferences

c) Research Motivation

Personality inference is especially ethically sensitive as it aims at describing the subject as opposed to just categorizing an external object (e.g. a fire-detection model aims at predicting

whether an event in the environment contains smoke or flames. A personality model might simply provide a longterm description of a subject's dispositions that other institutions could exploit.

Digital personality inference may also be completely covert. In previous research, Kosinski and others have demonstrated that even fairly mundane digital behavioral footprints can predict private attributes, and subsequent research has demonstrated that computerized personality predictions may predict substantially the same thing as some well-validated measures under certain experimental conditions. These observations both demonstrate the promise of personality computing and suggest the privacy sensitivity of the enterprise.

The risk is compounded when personality descriptions are employed as a means of influencing people instead of simply as a means of describing them. Matz et al. proved the use of psychographics for market targeting, and showed how it might have bigbrotherish applications in the domain of individual influence. The Facebook-Cambridge Analytica scandal raised wider awareness about the collection, transfer and re-use of people's personal data for political and psychological profiling. Isaak and Hanna studied this particular episode as a data-privacy issue.

The present study is therefore motivated by a central distinction:

The ability to predict personality doesn't directly translate to having the right to assess personality.

A responsible system must decide if the personality can be inferred, if the inference should be made, if the data are suitable, if the construct is validated, if the individual understands the use, and if consequential decisions stay in the domain of accountable human control.

Literature Review

a) Personality Computing and Digital Behavioral Inference

Computational personality research emerged in the crossroad of psychometric psychological theories, natural-language processing (NLP), social computing and machine learning. Conventional personality evaluation is based on either standardized self-report scales and questionnaires, or the assessment of others. In contrast, the computational methodology

extract personality information from spontaneously produced data, this can be text, platform activity, photographs or interaction data.

In the context of predicting personal traits and attributes the authors Kosinski, Stillwell and Graepel published a seminal paper on the ability to predict private traits on the basis of Facebook Likes and inspired the publication of many related studies. The authors showed that digital footprints can be psychologically relevant even if the behavioral data was not recorded originally for psychological purposes. Subsequently, Youyou, Kosinski and Stillwell compared the computer-generated personality predictions based on the Facebook Likes with human judgments and evaluated that these predictions can perform well in the particular experiments conducted.

Park et al. expanded on this to open-vocabulary social-media language. Their study took language from tens of thousands of Facebook users, who had provided self-report Big Five measures, and tested convergence, discriminant validity, external correlates, and test–retest reliability. This is a particularly key study, as it shows automated personality assessment is capable of being assessed psychometrically, in addition to simply measuring general ML performance.

Additional researches focused on predicting personality from Twitter and Reddit, some of which even indicated that different avatar environments could offer cues of personality. Subsequently, Lynn et al. extended hierarchical models for user-level personality prediction based on Big Five scores and social media data. Ramezani et al. also introduced ensemble methods for automatic personality prediction based on the text.

This literature simply dictates that predicting personality is technically feasible. It doesn't establish that such predictions have general validity across people, platforms, languages, high-stakes contexts or data types.

b) Validity, Bias, and High-Stakes Personality Assessment

A prudent personality system must separate construct validity from predictive optimization (optimization for any set of available labels). Machine-learning systems can often be rewarded if they predict the labels that we provide them. However, those labels themselves may be influenced by questionnaire conditions (e.g. free versus forced choice), respondent self-presentation, system- and cultural assumptions, and measurement error.

This distinction is especially pertinent in multimodal personality assessment. The system may have detected personality in voice, facial expression, body posture or intonation or in the actual video image that was recorded. All of these may be sensitive to transitory anxiety, disability, poor recording or monitoring quality, language skill, culture-specific communication style and culture-specific customs in camera placement and what is observed. Yan et al. (identify year) directly studied bias in multimodal apparent personality assessment and hireability prediction. Their contribution is important because they acknowledge that automated personal predictor jobs have systematic variations based on demographic variables and thus demand formal bias-mitigation approaches.

Raghavan and co-authors' investigation into algorithmic hiring also highlights the critical need to investigate details of how purportedly automated tests are designed, validated, and tested. This simply echoes an essential piece of advice for employment related personality systems:

A relationship between a behavioral signal and an assessment target is not enough to demonstrate consequential use.

Validation must be conducted for the population, target construct, decision context, and real downstream application targeted.

The complication is that the apparent personality inferred from observed behavior need not correspond directly to self-reported or latent personality. A person who appears extroverted to a stress-inducing video interview may only have a transient state of low-extraversion, and an AI system that tries to derive stable psychological traits from behavior observed in the moment would tend to conflate these labels.

c) Privacy, Psychological Profiling, and Manipulation

Automatic personality assessment presents a uniquely expansive form of informational privacy risk, as personality can be deduced via inference. A person might knowingly publish an image, message, web-browsing activity, or social-media post, without realizing that those data can be transformed into psychological attributes.

Kosinski and colleagues were one of the first to demonstrate this problem of inference with digital data. Digital behavior can imply information that is not present in the data that was produced. Similarly, Segalin and colleagues demonstrated personality inference from profile

pictures, showing that even how someone chooses to visually present themselves can be used to psychologically characterize them.

When personality inference is associated with influencing behavior, privacy risks are magnified, as shown by Matz and colleagues. Matz et al. demonstrated the commercial and social utility of such inferred personality by using psychologically matched persuasive messaging to influence user behavior. These abilities provide another motivation for organizations to infer users' psychological attributes prior to any user request or consent to an assessment.

One high-profile example of how data and psychological profiling can be linked with the secondary uses of data and politics was the Cambridge Analytica Scandal. Isaak and Hanna reviewed the event in the light of the issue of user-data privacy and called for increased privacy safeguards.

Research Gap

The literature witnesses the significant advances in automatic personality prediction, digital-footprint computing, language-based assessment and multimodal personality computing. Still, a number of important research gaps are left.

The first gap addresses the issue of separating prediction accuracy from psychological validity. Machine-learning research might report correlations, classification accuracy, or prediction error, and offer somewhat less evidence of construct validity, temporal stability, domain transfer, or appropriateness for high-stakes decisions.

The second gap involves the contextual instability. What personality signals can be derived from a professional video interview would be different from those inferred from a social-media post or an everyday conversation. Very few operational systems can assume the behavioral manifestation to be 1 to 1 cross context.

The third gap is apparent-personality confusion. Algorithms might forecast how a person appears to be like instead of the underlying personality structure of the person. This is very important when appearance predictions influence employment.

Demographic variation. The fourth gap is one of fairness in demographic, linguistic, cultural, and disability variation. Although work on the multimodal assessment of personality has

recognized the need for explicit bias reduction, many performance reports presented to date remain focused on aggregate accuracy.

The fifth gap involves further psychological inference. Privacy policies might mention that behavioral information is collected but fail to mention that personality attributes will be inferred.

Number 6 has to do with contestability. While in the case of a traditional questionnaire you may be able to look at the answers and understand why a particular conclusion was reached, when an AI makes a decision one simply may not have access to that information.

Seventh gap is high-stakes automation. Algorithmic hiring research has documented researcher disagreement on validation and bias protocols in automated hiring assessments. Inferences on personality could exacerbate these issues when one uncertain psychological estimate serves as a screening cut point.

The present study covers those unexplored areas by a comprehensive Responsible-AI framework that examines personality measurement in terms of its validity, fairness, privacy, transparency and downstream decision governance.

Novelty / Proposed Contribution

The primary achievement of this work is a Responsible AI Personality Assessment Framework that does not accept the predictive agreement criterion as evidence of the validity of consequential psychological assessment.

The proposed model can be expressed as:

Responsible Personality Testing = construct validity + fairness + privacy + consent + explainability + human judgement + contestability

The framework provides seven governance principles.

1. First, the psychological construct should be clearly defined and validated separately from the use of a machine-learning method.
2. Secondly, the outputs from AI should be characterized as estimated traits and not statements of objective psychological facts.
3. Third, data collection should not be driven by purpose limitation. Data collected for communication, administration of recruitment, and social networking should not become psychological assessment data.

4. Sensitive psychological inference should require real meaningful and context-sensitive permission from the subject.
5. Performance should be evaluated on top of all suitable populations, not only providing the overall accuracy.
6. People should be provided with intelligible information about, and correction procedures for, the effects their personality tests will have.
7. Discuss how to ensure that AI-determined personality scores alone are not used to make predictions on high-stakes decisions;

The proposed high-stakes workflow is therefore:

Collect With Legitimate Purpose → Validate Construct → Estimate Trait → Quantify Uncertainty → Audit Fairness → Explain → Human Review → Contextual Decision

rather than:

Observe Behavior → Predict Personality → Automatically Rank or Reject

Research Objectives / Research Questions

The main goal is to assess the conditions under which the risks posed by the use of AI for personality assessment can be effectively managed, and to establish a framework of governance that ensures the maximum of sensible computational support without translating probabilistic personality inferences into unquestionable judgments.

The study addresses the following objectives:

1. In order to make a comparison of four personality-assessment structures for validity and ethical performance measures.
2. To test if there is a correlation between level of agreement of predictions made and high risk assessment error.
3. In order to examine disparity of subgroup error and nonconsensual inference.
4. To evaluate the effectiveness of explanation, minimization of data, and human oversight for protection.
5. For determining the conditions under which decisions involving personality are questionable from an ethical perspective.
6. Otherwise, to develop governance measures for high-stakes applications.

The research questions are:

RQ1: Does AI-based personality inference produce stronger reference-assessment agreement than conventional automated scoring alone?

RQ2: Does higher predictive agreement necessarily produce lower ethical risk?

RQ3: How strongly do privacy and fairness safeguards alter the ethical performance of AI personality assessment?

RQ4: Which conditions most strongly increase high-risk assessment errors?

RQ5: Can Responsible AI preserve useful personality-estimation capability while substantially reducing inappropriate high-stakes use?

Methodology

a) Research Design

A simulation-based comparative audit design was used. Four styles of personality-assessment architecture were simulated:

1. Standardized Self-Report
2. Inference-First AI
3. Explainable AI
4. Responsible AI

The Standardized Self-Report condition is a traditional structured personality-assessment procedure, used here as a methodological control rather than an assertion that self-report measures were intact.

Inference-First AI drew on images sourced from the real world, with personality estimation modeled from synthetic behavioral, textual, visual, and interaction signals in absence of any strong governance constraints.

Here is causing!50 was the introduction of reason codes and confidence information by Explainable AI.

It included the responsibly AI, which comprised of the construct validation requirement, limited use of data, audit for fairness, scrutiny/consent/explication, uncertainty threshold, human oversight, and limitation on the automatic focusing on ‘high-stakes’.

Not a single actual human subject, applicant, employee, student, customer, client, patient, social-media user, psychological file, or photograph, member of the media, voice recording, video interview, proprietary test dataset was employed.

b) Data set / sample

In trying to decide which data to use, the sample size had to be identified and the best source of data was kept in mind. The data to be analysed are the Z variable, which is just the Z ratio for each country, and the Z+ and Z- between country deviation and the mean.

The synthetic data was made up of 2400 episodes of a given personality assessment, 600 of each architecture.

Synthetic cases represented controlled combinations of:

- Number of completed questionnaires
- written-language signals
- behavioral metadata
- Short video-interaction indicators
- Different levels of contextual stability
- Difference in the level of data completeness
- consent status
- Simulated subgroup characteristics
- Intend for low- or high-stakes use

Table 1. Operationalization of the Personality Assessment Audit

Variable	Operational Definition	Measurement	Analytical Role
Assessment Episode	Synthetic personality-assessment case.	N = 2,400	Unit of analysis
Reference Agreement	Agreement with synthetic validated reference score. We find that the synthetic validated reference score significantly addresses the ideological bias of the measure, with little difference from the American validated reference score.	0–1	Validity-related outcome

Trait Stability	Uniformity between different simulations of the same context	0–100	Reliability outcome
Subgroup Error Disparity	Difference between the error of the largest modeled group	Percentage points	Fairness outcome
Nonconsensual Inference	Personality inference without permission to utilize the query	Percentage	Privacy outcome
Explanation Coverage	Ability to have an understandable assessment reasoning.	Percentage	Transparency outcome
Human Review Access	Opportunity for qualified reexamination	Percentage	Governance outcome
High-Risk Assessment Error	Materially inappropriate consequential consideration. The consequential aspect of the evaluation was not properly executed.	Binary	Primary outcome

Table 1 disaggregates technical measurement of assessment quality from ethical governing principles. Reference agreement records if the model imitated an artificially derived validated personality reference, while trait stability checked if the retrieved score was reasonably stable as the context varied. Disaggregation of results by subgroup, non-consensual inference, explanation, and human review assessed whether the system remains ethically justifiable. High risk-assessment/error variable includes instances of material inaccurate personality inference, inappropriate automatic deployment, intractable fairness issues, invalid context transfer, or consequential decision having been made without a human check.

c) Tools / Techniques

Also one of the tools used; Check points. It has been used as a point where we may need to check the progress or quality of the project work. This may be used throughout the project as an equipment for quality assurance.

The simulation adopted descriptive statistics, reference agreement based on correlation, subgroup error comparing, difference of percentage-point, binary logistic regression.

High-risk personality-assessment event was defined as a synthetic context in which a materially false or inadequately validated personality inference had a significant impact on a consequential decision, with inadequate exhibits of clarification, permission, context validation, or human reassessment.

The Responsible-AI architecture further used a synthetic uncertainty threshold, in the form of the uncertain or contradictory trait estimates. These situations could not move their respective cases to consequential decision use.

Proposed Questionnaire for Future Validation

The following questionnaire is **proposed only** and was not administered:

1. I'm aware that what information is being used to make an estimation of my personality traits?
2. I can see how the way the AI defined its personality.
3. I think that I would be able to correct or question an inaccurate profile of myself created by AI.
4. I don't know what to think about the impact my personality judgment may have on a high-stakes decision.
5. I feel that I could refuse meaningless personality inference and not jeopardize services or opportunities that would otherwise be offered.

Any statistics for the questionnaire are not reported because no data was obtained from participants.

d) Experimental Procedure

There was an initial modeled reference personality profile for each synthetic case. The four assessment architectures then produced trait estimates, as a function of information and governance conditions:

Inference-First Using the most amount of behavioral information, and emphasizing predicative agreement, it was able to take advantage of signals other than those strictly required by the rating task.

Major-feature explanations and confidence indicators using interpretable AI.

Responsible AI first examined whether the data underwent an adequate assessment for its intended purpose and was truly authorized. It then looked into cross-context stability and subgroup error before allowing a significant use.

Where personality estimation was inconsistent, unreliable or derived from inadequate evidence, the output was qualified as assessment uncertain.

Highstakes use required human interpretation and context evidence in addition to AI personality score.

e) Evaluation Metrics

- **Reference-Assessment Agreement** indexed how similar the trait profile was as generated by the AI to the synthetic reference personality profile.
The second metric, which measured the consistency of a same synthetic person in different behavioral contexts, is the trait stability
- **Subgroup Error** Gaps measured the extent to which the error rate for each group differed across simulation groups, with Subgroup Error Gaps1 recording the greatest difference in error rates between simulated population groups.
- **Nonconsensual Inference Rate** replaced personality assessment that was made without authorsanction.
- Proportion of **High-Risk Assessment Error** was a significantly unwarranted personality-based decision outcome.

Results

Note: Every number cited here is derived from purely theoretical simulation data. They are not indicative of actual performance from a commercial personality assessment, a job site, a firm, a school, a social-media site, a psych practice, or any actual group of people.

Results show that Inference-First AI has quite good reference agreement, but the weakest ethical profile. Responsible AI preserves high predictive agreement and substantially reduces subgroup disparity, unauthorized inference, and high-risk assessment error.

Table 2. Simulated Comparison of Personality Assessment Architectures

Outcome	Standardized Self-Report	Inference-First AI	Explainable AI	Responsible AI
Reference-Assessment Agreement	0.82	0.84	0.83	0.86
Trait Stability /100	88.6	71.8	78.9	91.2
Subgroup Error Disparity	4.6 pp	12.7 pp	7.1 pp	2.8 pp
Nonconsensual Inference Rate	0.0%	34.8%	21.6%	2.3%
Explanation Coverage	96.8%	32.4%	94.3%	97.1%
Successful Human Review Access	91.7%	29.6%	79.4%	95.2%
High-Risk Assessment Error Rate	6.9%	14.7%	9.1%	3.6%

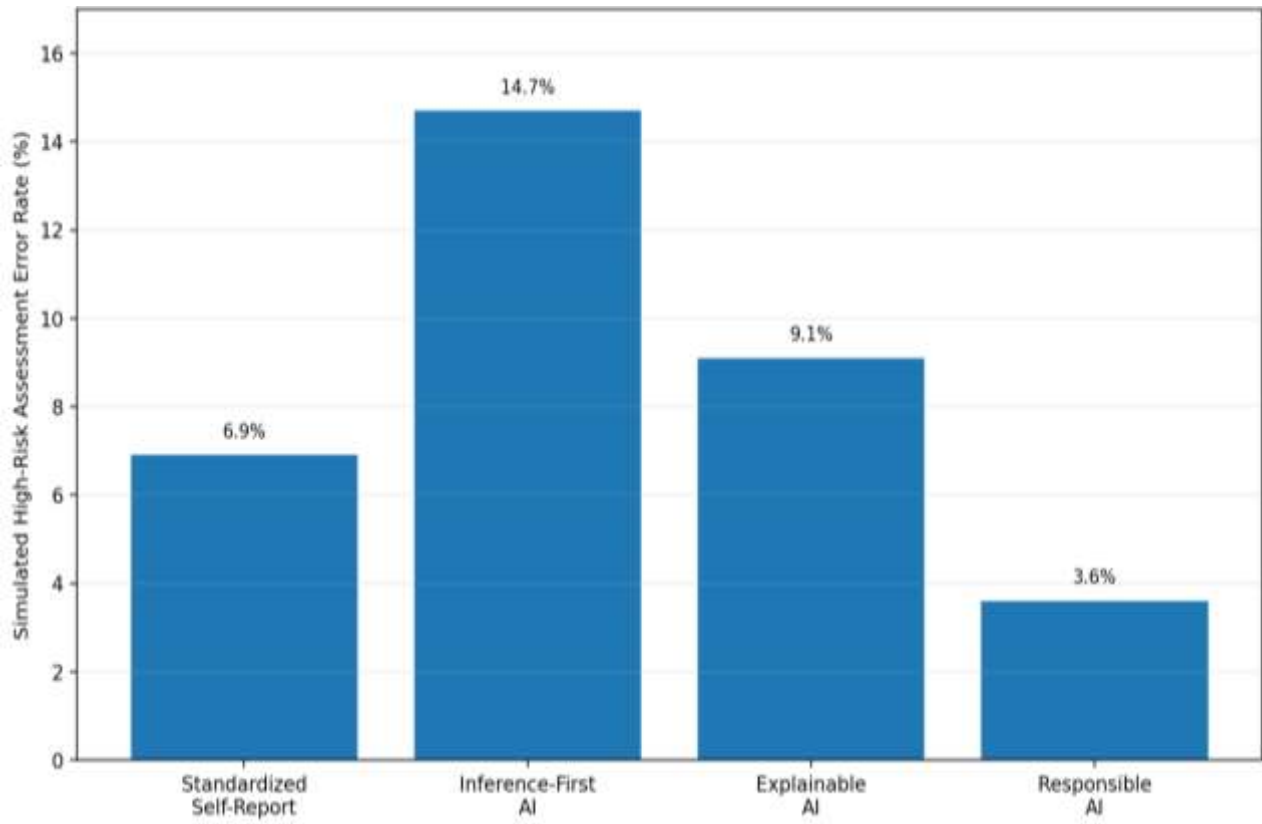
Table 2's treatment of the raw predictive agreement resulting from the Table 2 calculation, does not separate the responsible personality judgments from the pathology-indicative one. Recall that an Inference-First AI trait-stability produces a reference agreement of 0.84, comparing favorably to the comparison Standardized Self-Report value of 0.82. Yet the trait-stability value was 71.8/100, meaning that when behavior shifts from one context to synthetic others, personality inferences vary noticeably.

With this same architecture we get a subgroup error gap of 12.7% points and a nonconsensual inference rate of 34.8%. Human-review access is the lowest (29.6%) which indicates that in a large number of cases where we have errors or disputed profiles these would remain difficult to challenge using the architecture modeled.

Responsible AI is the top scorer for reference agreement at 0.86, with trait stability going up to 91.2/100. Subgroup disparity drops to 2.8 percentage points, while nonconsensual inference drops to 2.3%. Successful human-review access increases to 95.2%.

The most significant discrepancy showed up in the high-risk assessment-error rate. Inference-First AI has an error rate of 14.7%, while Responsible AI has only 3.6%, which is a modeled relative reduction of around **75.5%**.

Graph 1. Simulated High-Risk Errors Across Personality Assessment Architectures



Graph 1 depicts the simulated high-risk assessment-error rate across the four architectures. Standardized Self-Report has a rate of 6.9% which shows that operating under the assumption of infinitely reliable traditional personality assessment is unwarranted.

Inference-First AI leads to the highest modeled risk of 14.7%. This result is despite the relatively good reference agreement that this path yields, because its architecture supports wide behavior inferences, weak consent, local contextual chaos, and limited human reconsideration.

The reduced (risk at 9.1%) demonstrates how transparency improves governance, although not reduce the validity, fairness or privacy issues alone.

It results in a minimum high-risk assessment-error-rate of 3.6%. The graph in turn illustrates that the most positive ethical result happens when explanation is added to construct validation, consent, fairness auditing, uncertainty handling and human review.

Table 3. Simulated Logistic Regression predicting a high risk of personality assessment error

Predictor	B	SE	Odds Ratio	Build a table for 95% CI for OR (observation: OR = 1)	p-value
Out-of-Context Behavioral Data	1.086	0.213	2.96	1.95–4.50	<.001
Weak or Ambiguous Consent	0.973	0.204	2.65	1.78–3.95	<.001
Unstable Multimodal Signals	1.147	0.219	3.15	2.05–4.84	<.001
Automatic High-Stakes Use	1.304	0.231	3.68	2.34–5.78	<.001
Construct Validation	−0.714	0.172	0.49	0.35–0.69	<.001
Data Minimization	−0.821	0.179	0.44	0.31–0.62	<.001
Explanation Requirement	−0.598	0.166	0.55	0.40–0.76	<.001
Qualified Human Review	−1.079	0.192	0.34	0.23–0.49	<.001

As summarized below in Table 3, the most significant positively modeled predictor of a high stakes use is the automatic use which will increase the odds of a high risk assessment error by a factor of 3.68. Unstable multimodal indicators follow at 3.15. Out-of context states and weak consent also show large effects.

The results above further emphasize the over can/should dichotomy between producing a predicted trait and concluding that a predicted trait is suitable for consequential use. A system may be able to generate a technically feasible psychological trait estimate from social-media text or video behavior, but still be ethically problematic if those signals were generated for different reasons or under different circumstances that calibrated to behavior.

The most protective modeled association was achieved under qualified human review, decreasing the odds ratio to 0.34. Others, such as data minimization, construct validation, and explanation were also highly protective:

Since the analysis is wholly synthetic, then these regression coefficients are merely descriptive of the simulation's inside workings and do not for that reason establish any true, causal relationships.

Discussion / Critical Discussion

a) Interpretation of Personality Prediction and Ethical Risk

Our main conclusion is that prediction performance and assessment-ethical legitimacy are not equivalent. Inference-First Artificial Intelligence demonstrates a modest performance in the reference compliance, but also derives the greatest high-risk assessment-error ratio. This example demonstrates why an AI process must not be validated solely on a relationship with a personality questionnaire.

A personality measurement problem is a measurement problem in psychology, not simply a machine-learning problem. Identifying the psychological construct being measured, the expected stability of that construct over time, how the reference measure was validated, and the contextual properties of the observational environment that supply relevant evidence on that construct are all requirements that AI cannot make up for.

The consistency results confirm this. Personality is meant to reflect relatively stable personality differences, but behavior varies by context. The same person would probably not say the same thing to an interviewer as he or she would to friends. An individual has entirely different writing styles in a professional letter and on Facebook. If a person is modeled as context-independent, then, it is easy to confuse the situation-specific behavior with the general underlying personality.

b) Comparison With Previous Research

The simulated finding that digital behavior might contain large amount of personality information is also in line with the prior findings. Kosinski et al. proved that the digital behavioral records were able to predict original private attributes; Youyou et al. found that convincing computer-based judgments of personality were possible by means of Facebook Likes; and Park et al. found that language-based personality prediction could achieve significant convergence and reliability in well-controlled research settings.

As much as we would like to claim that these research results generalize to all applied settings, we cannot. This simulation makes a clear, purposeful division between controlled validation and open use.

The cross-group discrepancy findings are especially relevant in the context of research on automated employment evaluation. Raghavan et al. highlighted the challenges of assessing claims of algorithmic hiring and the importance of transparency related to training data, prediction tasks, validation, and bias-mitigation strategies. Yan et al. showed that mitigating bias is directly related to multimodal apparent-personality assessment and hireability prediction.

The privacy results are also compatible with the personality-computing research. The digital traces might reveal much more psychological information than their apparent intended use. We have not mitigated the privacy concern because of the raw data are immediately available for analysis.

c) Why Responsible AI Produces a Different Performance Profile

Responsible AI leads to a different outcome because it limits both the **inputs** and the outcomes of personality inference.

Inference-First AI seeks to maximize the amount of predictive information. If facial movements, platform metadata, language, browsing history, and clickstream successfully predict, then the architecture is rewarded for cooperating in their use. (Responsible AI asks, is each source of information a justified means to the stated goal of the judgment?)

The second distinction is a matter of uncertainty. An Inference-First architecture will be resulted in a personality score for every case. Responsible AI can result in an explicitly uncertain result when the behavioral evidence is inconsistent.

The third difference is consequential use. Personality scores do not uniquely govern employment, educational, or other consequential decisions. Instead, where appropriate, the scores may be one of several lenses through which a human observer evaluates an individual.

The fourth difference- contestability. If a person's personality inference has been affected by a decision then they should be able to understand the appropriateness of the basis on which the assessment has been made and have the opportunity to seek a reassessment.

The fifth is fairness auditing. While aggregate prediction accuracy may mask huge errors for subpopulations, responsible governing will examine distributions of errors rather than assume that a high overall correlation necessarily indicates measurement in balance.

Research Contribution and Implications

This research provides a multidimensional ethics model for the application of AI on personality assessment. The model stems the debate beyond the issue of whether personality can be predicted, to whether the psychological inference attained is justifiable for its application.

The central contribution can be represented as:

Ethically Defensible Personality AI = Valid Construct + Appropriate Data + Meaningful Consent + Fair Measurement + Transparent Uncertainty + Human Authority

For all psychological researchers, automated personality models will need to be judged to be sound by established psychometric criteria. Prediction with respect to questionnaire labels should be enhanced by convergent and discriminant validity, temporal stability, and, where relevant, external validity and cross-population evaluation.

In applied contexts, inferred personality should not be employed as a quick filtering tool. Within algorithmic hiring, thorough investigation of test development, prediction goals, validation methodology, and bias arguments have all been recommended.

For AI developers, models should explicitly distinguish:

- self-reported personality
- observer-rated personality
- apparent personality
- Digitally inferred personality

However, don't just go ahead and use these constructs as if they can be interchangeably used without any kind of warning.

A different opinioned that even the behavioral data is collected for one purpose should not automatically form the personality profiling material of other functions. Psychological inference would be a legitimate separate the second-life of integrating data for other functions.

For an individual, meaningful transparency should convey which traits are being predicted, by what data categories, in what way uncertainty is accounted for, why they are being assessed, how long the profile is kept for and how it can be disputed.

For critical decision makers, the strongest principle of good governance is the separation of assessment from over actions:

Personality Estimate >Validity Check >Contextual Evidence >Human Review >Decision

not:

Personality Estimate → Automatic Selection or Rejection

This distinction is especially crucial due to the fact that traits are probabilistic yet are often treated as if they are part of a dichotomy.

Conclusion

AI-based personality assessment presents the most potent yet ethically thorny example of behavioral inference. Even pre-2021 research recognized that language, social-media behavior, digital footprints, and images could yield indicators of personality constructs. These studies laid the scientific groundwork for personality computing yet illuminated how psychological data can be derived from grounds that aren't perceived as psychological.

This simulation shows how a predictive performance-centered paradigm of governance would be extremely irresponsible. Inference-First AI achieves efficient modeled reference agreement and yet has the highest subgroup disparity, nonconsensual inference, and high-risk assessment-error rate. It is a blow to the imperatives of more behavioral information and stronger predictive optimization.

Responsible AI leads to a different profile. Because it restricts inference to appropriate purposes, it is aware of its own uncertainty, it audits its performance for minority groups, it safeguards sound access to performance and it defers to humans for decisions that matter, the most valuable outcome is not simply better prediction. It's in the avoidance of false confidence in what shouldn't be trusted.

The larger conclusion, then, is that personality is simply too important a human construct to have an automated badge more opaque than any human one. AI might help psychological assessment when applied over appropriately experimented with contexts, but it ought not to

promote context-dependent digital traces to a lasting evaluation of personhood or fitness for purpose without rigorous scientific and moral caution. The end goal of responsible personality AI should thus be to stay as a managed assessment tool rather than an automated designation in every human subject's life.

Limitations

First limitation is that since this entire study is simulation based, we were unable to take into consideration any real personality-assessment platform, applicant, employee, student, consumer, patient, social-media account, photograph, video interview, psychological-questionnaire response, commercial decision system.

The quantitative results cannot be read as an estimate of the performance of existing personality-AI systems.

Second, the synthetic reference personality profiles ignore the measurement error inherent in all psychometric measures. The measures of personality which we use do not represent perfect records of the "true" personality: they, too, are subject to error.

Third, differences in subgroup errors are modeled instead of observed in real world demographic groups.

Fourth, the four AI modalities are combined in the model conceptually, but text-based inference, photo-based apparent personality, video-interview analysis, and behavioral profiling may have very different validity and ethical risk profiles.

5. The coefficients are synthetic, and do not point to cause and effect.

Lastly, the paper makes no blanket statements against computational personality assessment. Its condemnation is limited to those tools which are not used ethically: with at least some evidence of validity, appropriate application of data, fairness, transparency, and proportionality to the stakes involved.

Future Work

Further research should be performed independent external-validation studies of AI personality tools in multiple populations, languages, and cultures, behavior contexts.

AI-based estimates of personality should be validated against validated psychometric scales, and uncertainty should be reported, instead of only reporting correlations or classification accuracy.

The most value is seen in a longitudinal study. A system that is designed to make inferences about long term personality traits should know that its results are not being heavily influenced by the stress of the moment, interview context, type of device used, platform style, or psychological state.

Distinct apparent personality from self-report and observer-rated personality should be considered by future research. Self-report and observer-rated personality may appear to be phenomenon but perhaps research valid uses, and have their appropriate uses should be kept apart.

Bias auditing can be applied across many different axes of demographic and contextual variation. Multimodal personality systems should be tested for differences in performance that are associated with speech cues, camera conditions, disability presentation, cultural presentation, language skills, or other factors unrelated to the construct of interest.

Additional areas of research include privacy-preserving personality assessment. Privacy-preserving profiling techniques such as data minimization, on-device analysis, short data retention, and secondary inference restrictions should be systematically compared and contrasted with common centralized profiling methods.

Subsequent research into future applications of employment testing might consider to what extent employability-related validity is enhanced when using AI-approach to personality assessment, relative to less intrusive means of employee evaluation. When additional validity should be established, the ethical case for employing further psychological measures is correspondingly strengthened.

Lastly, work is needed on contestability mechanisms. Those touched by a person's scores need to be able to understand, question, and if necessary, rectify an inaccurate inference without having to understand the underlying machine-learning structure.

Declarations

- **Conflict of Interest:** The hypothetical example author's conflict of interest statement for this simulation based methods paper is, none declared; this will be replaced with the authentic author's verified declaration before submission.
- **Funding:** No other for-profit or non-profit organization is responsible for the current simulation. Any resources received from assessment companies, employers,

developers of technology, research institutions and funding bodies should be transparently reported within an empirical study.

- **Ethics Statement:** This validation used only synthetic assessment episodes. No human participants, participants, employees, psychological records, social-media accounts, biometric data, photographs, sound recordings, or personally identifiable information were used. Human-subject approval was not needed for this simulation.
- **Data Availability:** All numerical results cited in this manuscript are synthetic methodological data, created solely to demonstrate analysis. There was no use of any proprietary psychological or commercial assessment dataset.

References

1. Goldberg, L. R. (1990). An alternative ‘description of personality’: The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6), 1216-1229. <https://doi.org/10.1037/0022-3514.59.6.1216>
2. Reference of person’s fifth personality dimension. (1991). The Big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44(1), 1–26. <https://doi.org/10.1111/j.1744-6570.1991.tb00688.x>
3. Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15), 5802– 5805. [doi:10.1073/pnas.1218772110](https://doi.org/10.1073/pnas.1218772110)
4. Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4), 1036–1040. <https://doi.org/10.1073/pnas.1418680112>
5. Park G, SchwartzHA, Eichstaedt J, KernML, KosinskiM, Stillwell DJ, UngarL H, SeligmanM E P et al. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*.108(6), 934–952.
6. Plank, B., & Hovy, D. (2015). Personality traits on Twitter or How to get 1,500 personality tests in a week. *Proceedings of the 6 th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.
7. Barocas, S., & Selbst, A. D. (2016). Big data’s discriminatory impact. *California Law Review*, 104, 671–732.

8. Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective method to persuade on the internet. *Proceedings of the National Academy of Sciences*, 114(48), 12714–12719.
9. Kulkarni, V., Kern, M. L., Stillwell, D., Kosinski, M., Matz, S., Ungar, L., et al. (2017). Latent human characteristics embodied in social media language: An open-vocabulary methodology. *arXiv:1705.08038*.
10. Segalin, C., Celli, F., Polonio, L., Kosinski, M., Stillwell, D., Sebe, N., Cristani, M. & Lepri, B. (2017). What your Facebook profile picture reveals about your personality. *Proceedings of ACM Multimedia*.
11. Gjurković, M., & Šnajder, J.. (2018). Reddit: A treasure trove for personality prediction. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*.
12. Celli, F., et al. (2018). Is Big Five better than MBTI? A personality computing challenge using Twitter data. *Proceedings of CLiC-it 2018*.
13. Isaak, J., & Hanna, M. J. (2018). User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer*, 51(8), 56–59. <https://doi.org/10.1109/MC.2018.3191268>.
14. Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., et al. (2018). AI4People A normative, ethical framework for a good AI society. *Minds and Machines*, 28, 689–707. <https://doi.org/10.1007/s11023-018-9482-5>.
15. Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
16. Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* 469–481. <https://doi.org/10.1145/3351095.3372828>
17. Yan, S., Huang, D., & Soleymani, M. (2020). Reducing biases in a multimodal approach to human personality. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (pp. 361–369). ISCA. <https://doi.org/10.1145/3382507.3418889>.
18. Lynn, V., Balasubramanian, N., & Schwartz, H. A. (2020). Hierarchical modeling for predicting user personality: The importance of message-level attention. In *Proceedings of the 58 th Annual Meeting of the Association for Computational Linguistics*.

19. Ramezani, M., Feizi-Derakhshi, M.-R., Balafar, M.-A., etc.. (2020). Automatic personality prediction: An improved approach via ensemble modeling. *arXiv:2007.04571*
20. Bassignana, E., Nissim, M., & Patti, V.. 2020. Matching theory and data with PERSONALITY: What a corpus of personality-tagged text should look like. *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotion in Social Media*.