



An external stability audit framework to test the validity of personality prediction in AI hiring

Alene K. Rhea^{1,2} · Kelsey Markey^{1,2} · Lauren D'Arinzo^{1,2,3} ·
Hilke Schellmann⁴ · Mona Sloane² · Paul Squires⁵ · Falaah Arif Khan^{1,2} ·
Julia Stoyanovich^{1,2,6}

Received: 6 October 2021 / Accepted: 5 August 2022
© The Author(s) 2022

Abstract

Automated hiring systems are among the fastest-developing of all high-stakes AI systems. Among these are algorithmic personality tests that use insights from psychometric testing, and promise to surface personality traits indicative of future success based on job seekers' resumes or social media profiles. We interrogate the validity of such systems using stability of the outputs they produce, noting that reliability is a necessary, but not a sufficient, condition for validity. Crucially, rather than challenging or affirming the assumptions made in psychometric testing — that personality is a meaningful and measurable construct, and that personality traits are indicative of future success on the job — we frame our audit methodology around testing the underlying assumptions made by the vendors of the algorithmic personality tests themselves. Our main contribution is the development of a socio-technical framework for auditing the stability of algorithmic systems. This contribution is supplemented with an open-source software library that implements the technical components of the audit, and can be used to conduct similar stability audits of algorithmic systems. We instantiate our framework with the audit of two real-world personality prediction systems, namely, Humantic AI and Crystal. The application of our audit framework demonstrates that both these systems show substantial instability with respect to key facets of measurement, and hence cannot be considered valid testing instruments.

Keywords Algorithm Audit · Validity · Stability · Reliability · Hiring · Personality

Responsible editor: Toon Calders.

L. D'Arinzo: The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions, or viewpoints expressed by the author. Approved for Public Release; Distribution Unlimited. Public Release Case Number 21-3193. © 2021 The MITRE Corporation. All rights reserved.

✉ Julia Stoyanovich
stoyanovich@nyu.edu

Extended author information available on the last page of the article

1 Introduction

AI-based automated hiring systems are seeing ever broader use and have become as varied as the traditional hiring practices they augment or replace. These systems include candidate sourcing and resume screening to help employers identify promising applicants, video and voice analysis to facilitate the interview process, and algorithmic personality assessments that purport to surface personality traits indicative of future success. HireVue, a company that sells one of these systems, estimates that the “pre-hire assessment” market is worth \$3 billion annually (Kelly-Lyth 2020). Indeed, most Fortune 500 companies are using some form of algorithmic hiring (Schellmann et al. 2021a). Ian Siegel, the CEO of ZipRecruiter (a popular online employment marketplace), estimates that 75%–100% of all submitted resumes are now read by software, and that only a small fraction of those go on to be read by humans (Schellmann et al. 2021a).

In this paper, we focus on automated pre-hire assessment systems, as some of the fastest-developing of all high-stakes uses of AI (Kelly-Lyth 2020). The popularity of automated hiring systems in general, and of pre-hire assessment in particular, is due in no small part to the hiring sector’s collective quest for efficiency. Employers choose to use them to source and screen candidates faster and with less paperwork and, in a world reshaped by the COVID-19 pandemic, with as little in-person contact as is practical. Job seekers are, in turn, promised a more streamlined job search experience, although they rarely have a choice in whether they are screened by an automated system, and they are typically not notified when algorithmic screening is used (Stoyanovich 2021). The flip side of efficiency potentially afforded by automation is that job seekers, the general public, and even employers themselves rarely understand how these systems work and, indeed, whether they work. Is a resume screener identifying promising candidates or is it picking up irrelevant—or even discriminatory—patterns from historical data, potentially exposing the employer to legal liability? *Are job seekers participating in a fair competition if they are systematically unable to pass an online personality test, despite being well-qualified for the job* (Weber and Dwoskin 2014)?

Personnel selection is an especially sensitive, high-stakes application of AI. Hiring decisions are often of great consequence to the financial and emotional well-being of the job seekers (Bendick 2007), and in aggregate contribute to widespread economic inequality (Blau et al. 2013; Hegewisch et al. 2010). Consequences for hiring organizations can be substantial as well: if their selection procedures are arbitrary or unfair, they risk litigation and class action lawsuits. As such, any algorithms deployed in the field of hiring deserve rigorous scrutiny.

This realization is starting to be codified in laws and regulation. An important recent example is Local Law 144 of 2021 that requires bias auditing of “automated employment decision tools” used by employers in New York City, and also mandates disclosure about the use of these tools to job seekers before they are screened (New York City Council 2021). Another example is the Artificial Intelligence Act (AI Act), proposed by the European Commission in 2021 to serve as a common regulatory and legal framework for AI in the European Union (The European Commission 2021). The Act states that “AI systems used in employment, workers management and access to self-employment, notably for the recruitment and selection of persons, for making

decisions on promotion and termination and for task allocation, monitoring or evaluation of persons in work-related contractual relationships, should also be classified as high-risk, since those systems may appreciably impact future career prospects and livelihoods of these persons,” and subjects such systems to strict oversight requirements.

Reports of algorithmic hiring systems acting in ways that are discriminatory or unreliable abound (Bandy 2021; Bogen and Rieke 2018; Dastin 2018; Datta et al. 2015; Köchling and Wehner 2020; Stark and Hutson 2021). In a recent example, when testing automated phone interview software, Hilke Schellmann found that the system produced “English competency” scores even when the candidate spoke exclusively in German or Chinese (Schellmann et al. 2021b). This finding undermines the *validity* of the tool, and crystallizes the fact that black-box algorithms may not act as we expect them to.

In our work we interrogate the validity of algorithmic pre-hiring assessment systems of a particular kind: those that purport to estimate a job seeker’s personality based on their resume or social media profile. Our focus on these systems is warranted both because the science behind personality testing (algorithmic or not) in hiring is controversial (Emre 2018; Lussier 2018; Sloane 2021), and because algorithmic personality tests are rarely validated by third-parties (Schellmann et al. 2021a). Warning against this trend, Chamorro–Premuzic *et al.* (Chamorro-Premuzic et al. 2016) write in the *Journal of Industrial and Organizational Psychology*: “shiny new talent identification objects often bamboozle recruiters and talent acquisition professionals with no regard for predictive validity.” Despite this warning, unvalidated use of these “objects” continues. For example, as we will discuss in Sect. 4, DiSC, a psychometric instrument used by several algorithmic personality assessment systems, has not been validated in the hiring domain, and the company that produces DiSC specifically warns against using it for pre-employment screening.

In our work, we focus on *stability*, by which we refer to a property of an algorithmic system whereby small changes in the input lead to small changes in the output, noting that this property is a necessary, albeit not a sufficient, condition for validity. Our approach is to (1) develop a methodology for an *external audit of the stability* of algorithmic personality predictors, and (2) instantiate this methodology in an audit of two real-world systems, *Humantic AI* and *Crystal*. Crucially, based on the insights of Sloane *et al.* (Sloane et al. 2022), we frame our methodology around *testing the underlying assumptions made by the vendors of the algorithmic personality tests themselves*.

Humantic AI and *Crystal* were selected as audit subjects because they each produce quantitative personality traits as output, accept easily-manipulated textual features as input, and allow multiple input types. These systems also have substantial presence in the algorithmic hiring market: *Humantic AI* reports that it is used by Apple, PayPal and McKinsey,¹ and *Crystal* claims that 90% of Fortune 500 companies use their products, though neither company distinguishes between use for hiring and use for other purposes, such as sales.²

¹ <https://humantic.ai/>.

² <https://www.crystalknows.com/>.

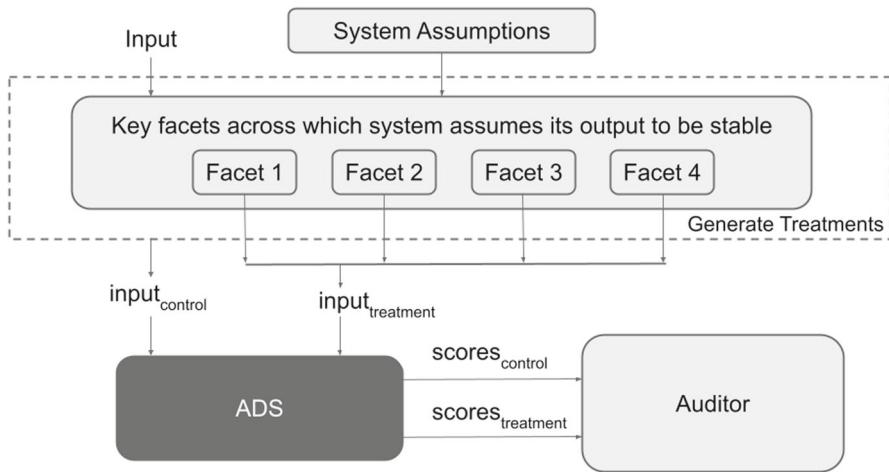


Fig. 1 Socio-technical framework for stability auditing, discussed in detail in Sect. 3

In this paper, we make the following contributions:

1. We provide an overview of the key literature on psychometric testing applied to hiring and on algorithm auditing with a particular focus on hiring (Sect. 2). We find that reliability is seen as a crucial aspect of the validity of a psychometric instrument, yet it has not received substantial treatment in algorithm audits.
2. We propose a socio-technical framework for auditing the stability of algorithmic systems (Sect. 3). Figure 1 gives an overview of our proposed methodology. As part of this contribution, we develop an open-source software library that implements the technical components of the audit, and can be used in stability audits of automated decision systems (ADS), with suitable input data, treatment generation techniques, and choice of stability metrics. Our library can be extended with additional input and output data types, treatment and control generation methods, and choice of stability metrics.
3. We instantiate this methodology in an external stability audit of *Humantic AI* and *Crystal*, two black-box algorithms that predict personality for use in hiring, over a dataset of job applicant profiles collected through an IRB-approved study (Sect. 4). The application of our audit framework surfaces substantial instability with respect to important facets of measurement in both these systems, results are presented in Sect. 5. For example, we find that personality profiles returned by both *Humantic AI* and *Crystal* are substantially different depending on whether they were computed based on a resume or a LinkedIn profile, violating the assumption that an algorithmic personality test is stable across input sources that are treated as interchangeable by the vendor. Further, *Crystal* frequently computes different personality scores if the same resume is given in PDF vs. in raw text format, violating the assumption that the output of an algorithmic personality test is stable across job-irrelevant variations in the input.

We discuss the results and limitation of our work in Sect. 6, and conclude in Sect. 7.

2 Background and related work

2.1 Validity and reliability in psychometric theory applied to hiring

Personality testing in hiring. Since the early 1900s, personnel selection practices have relied on the use of psychometric instruments such as personality tests to identify promising candidates (Scroggins et al. 2008), and the use of these tests continues to be wide-spread (Meinert 2015). And although this practice is both longstanding and wide-spread, it has been met with skepticism from industrial-organizational (I-O) psychologists due to validity and reliability concerns, and even led to disagreements about whether personality itself is a meaningful and measurable construct (Scroggins et al. 2008). A comprehensive literature review of personality testing in personnel selection published in 1965 found little evidence of predictive validity, and concluded that “it is difficult to advocate, with a clear conscience, the use of personality measures in most situations as a basis for making employment decisions” (Guion and Gottier 1965). Several other surveys would come to the same conclusion in the following decades (Hough et al. 1990; Schmitt et al. 1984), yet, HR professionals continued to use personality testing for hiring (Scroggins et al. 2008). The rise of the “Big Five” model of personality in the 1990s led to wider acceptance of personality testing in hiring amongst I-O psychologists, albeit not without controversy. (See Sect. 4.1 for more on the Big Five.)

The use of a traditional personality test in personnel selection relies on the following assumptions:

- The personality traits being measured are meaningful constructs;
- The test is a valid measurement instrument: it measures the traits it purports to measure;
- The test is a valid hiring instrument: its results are predictive of employee performance.

Validity and reliability of psychometric instruments. Within the field of psychometrics, instruments are considered useful only if they are both reliable and valid (Cardinet et al. 1976; Carmines and Zeller 1979). *Reliability* refers to the consistency of an instrument’s measurements, and *validity* is the extent to which the instrument measures what it purports to measure (Mueller and Knapp 2018). Reliability is a necessary (although not a sufficient) condition for validity (Nunnally and Bernstein 1994). Thus, when considering psychometric instruments, the question of reliability is central to the question of validity.

Reliability can be measured across time (*test-retest reliability*), across equivalent forms of a test (*parallel forms reliability*), across testing environment (*cross-situational consistency*), etc. (Mueller and Knapp 2018). Each of the dimensions across which measurements are compared is referred to as a *facet*, such that we can talk about reliability with respect to some facet (e.g., time) that varies between measurements, while other facets (e.g., test location) are held constant (Cardinet et al. 1976). Under Classical Test Theory (CTT), measurements can be decomposed into a true score and a measurement error (Schmidt et al. 2003). The true score is the value of the

underlying construct of interest (e.g., extraversion). Measurement error can be broken down across various experiment facets (Schmidt et al. 2003).

Reliability is usually measured and evaluated with correlations. Although 0.80 is often cited as an acceptable threshold of reliability, Nunnally and Bernstein differentiate between standards used to compare groups (for which 0.80 is an appropriate reliability), and those used to make decisions about individuals. For the latter type of test, they advise that 0.90 should be the “bare minimum,” and that 0.95 should be the “desirable standard” (Nunnally and Bernstein 1994).

Algorithmic personality tests, on which we focus in this paper, constitute a category of psychometric instruments, and are thus relying on the same assumptions—about test validity as a measurement instrument and as a hiring instrument—as do their traditional counterparts. Guzzo *et al.* caution that reliability and validity are “often overlooked yet critically important” in big-data applications of I–O psychology (Guzzo et al. 2015). In our work, we aim to fill this gap by interrogating the reliability of algorithmic personality predictors. Because the objects of our study are algorithmic systems that are used by employers in their talent acquisition pipelines, our work falls within the domain of hiring algorithm audits, discussed next.

2.2 Auditing of hiring algorithms

Background on algorithm auditing. The algorithm audit is a crucial mechanism for ensuring that AI-supported decisions are *fair, safe, ethical*, and correct. Increasing demand for such audits has led to the emergence of a new industry, termed Auditing and Assurance of Algorithms by Koshiyama *et al.* (Koshiyama et al. 2021).

Scholarly work on algorithm auditing acknowledges that auditing frameworks are inconsistent in terms of scope, methodology, and metrics (Bandy 2021; Brown et al. 2021; Koshiyama et al. 2021; Raji et al. 2020). In this landscape that offers many frameworks, yet minimal technical guidance, auditors are left to define their own scope. As argued by several authors, stakeholder interests should be central to the task of scoping (Brown et al. 2021; Fjeld et al. 2020; Metcalf et al. 2021; ORCAA 2020; Raghavan et al. 2020; Raji et al. 2020; Razavi et al. 2021; Sloane et al. 2022; Vecchione et al. 2021). Sloane *et al.* argue that audits ought to be specific to the domain and to the tool under study (Sloane et al. 2022).

In the United States, much of the audit literature surrounding predictive hiring technology is concerned with legal liability as laid out in the Uniform Guidelines on Employee Selection Procedures (UGESP) (Kim 2017; Raghavan et al. 2020; Wilson et al. 2021). These guidelines, adopted by the Equal Employment Opportunity Commission in 1978, revolve around a form of discrimination called disparate impact, wherein a practice adversely affects a protected group of people at higher rates than privileged groups

Equal Employment Opportunity Commission (EEOC) et al. (1978). As a result, audits of AI hiring systems are often specifically concerned with adverse impact (Chen et al. 2018; ORCAA 2020; Wilson et al. 2021). It is often noted that avoiding liability is not actually sufficient to ensure an ethical system; that is, a lack of adverse impact should

be a baseline rather than the goal (Barocas and Selbst 2016; ORCAA 2020; Raghavan et al. 2020; Wilson et al. 2021).

The main contribution of our work is a socio-technical audit methodology developed to measure the stability of personality prediction systems used in the hiring domain, and an open-source library that generalizes the technical components of this framework for use more broadly in stability auditing. We further instantiate this framework on two real-world personality prediction systems. As we will discuss in Sect. 3, we build on Sloane *et al.* (Sloane et al. 2022) to interrogate the assumptions encoded by these systems.

Treatment of reliability in algorithm audits The audit literature is inconsistent in whether reliability is included as a concern and, if it is, how it is defined and treated. Several impactful lines of work do not consider reliability (Hagendorff 2020; Langenkamp et al. 2020; Metcalf et al. 2021; Sandvig et al. 2014; Sühr et al. 2021; Venkatadri et al. 2018; Wilson et al. 2021). Of the works that do take reliability under consideration, some refer to this concept as *stability* (Brown et al. 2021; Koshiyama et al. 2021; Robertson et al. 2018; Sloane et al. 2022; Riksrevisjonen 2020), others as *reliability* (Fjeld et al. 2020; Mökander et al. 2021; Raji et al. 2020; Riksrevisjonen 2020; Shneiderman 2020), and others yet as *robustness* (Chen et al. 2018; Fjeld et al. 2020; Mökander et al. 2021; Oala et al. 2020; ORCAA 2020). Bandy Bandy (2021) forgoes specific terminology and simply refers to changes to input and output. This difference in treatment is more than terminological: stability relates to local numerical analyses, whereas robustness tends to refer to broad, system-wide imperviousness to adversarial attack, and reliability connotes consistency and trustworthiness.

This inconsistency is part of a larger problem within sensitivity analysis—the formal study of how system inputs are related to system outputs. Razavi *et al.* observe that sensitivity analysis is not a unified discipline, but is instead spread across many fields, journals and conferences, and notes that lack of common terminology remains a barrier to unification (Razavi et al. 2021). In our work, we use the term *stability* to refer to a property of an algorithm whereby small changes in the input lead to small changes in the output. We adopt a psychometric definition of *reliability*, which we use to guide the way in which we measure stability. By considering algorithms within their socio-technical context, we can also translate between numerical stability and broader *robustness*.

Although reliability has not been centered in algorithm audits, the importance of model stability has long been established (Turney 1995). The 2020 manifesto on responsible modeling by Saltelli *et al.* (Saltelli et al. 2020) underscores the importance of sensitivity analysis, and both the European Commission European Commission (2021) and the European Science Academies Science Advice for Policy by European Academies (SAPEA) (2019) have called for sensitivity auditing in the policy domain. As detailed by Razavi *et al.*, sensitivity audits have also been applied in the domains of education (Araujo et al. 2017), food security (Saltelli and Lo Piano 2017), public health Lo Piano and Robinson (2019), and sustainability (Galli et al. 2016), see (Razavi et al. 2021). We argue that algorithm auditors should consider stability among the critical metrics they select from, as suggested by Brown *et al.* (Brown et al. 2021).

Our work is synergistic with two recent lines of work that contribute substantive quantitative methodologies for auditing algorithm stability. The first, (Xue et al. 2020),

introduces a suite of tools to study individual fairness in black-box models, while the second, (Sharma et al. 2020), offers a unified counterfactual framework to measure bias and robustness. Sharma *et al.*'s methodology relies on access to the features being used by the model, whereas the methods proposed by Xue *et al.* and by our work only require query access to black-box models. The key distinction between Xue *et al.* and our work is that Xue *et al.* build on notions of individual fairness that can be encoded by Wasserstein distance, while we approach stability through a socio-technical lens, borrowing metrics that are familiar to I–O psychologists.

Audit scope. A number of recent algorithm audits focus on tools used at various stages in hiring pipelines. Wilson *et al.* (Wilson et al. 2021) and O'Neil Risk Consulting and Algorithmic Auditing (ORCAA) (ORCAA 2020) each focus on tools for pre-employment assessment (i.e., candidate screening). Raghavan *et al.* (Raghavan et al. 2020) evaluate the public claims about bias made by the vendors of 18 such tools. Chen *et al.* (Chen et al. 2018) audit three resume search engines, Hannák *et al.* (Hannák et al. 2017) audit two online freelance marketplaces, and De-Arteaga *et al.* (De-Arteaga et al. 2019) builds and evaluates several classifiers that predict occupation from online bios. All of these studies focus primarily on bias and discrimination. It is also common to frame these audits around the promises made by the companies in their public statements (ORCAA 2020; Raghavan et al. 2020; Wilson et al. 2021). By contrast, in our work we focus on auditing stability, which is a necessary condition for the validity of an algorithmic hiring tool.

Access level is a critical factor in determining audit scope. Audits can be internal (where auditors are employed by the company being audited), cooperative (a collaboration between internal and external stakeholders), or external (where auditors are fully independent and do not work directly with vendors). Sloane *et al.* (Sloane et al. 2022) explain that the credibility of internal audits must be questioned, because it is advantageous to the company if they perform well in the audit. Ajunwa Ajunwa (2021) argues for both internal and external auditing imperatives, with the latter ideally performed by a new certifying authority. Brown *et al.* Brown et al. (2021) offer a flexible framework for external audits that centers on stakeholder interests. Bogen and Rieke Bogen and Rieke (2018) stress the importance of independent algorithm evaluations and place the burden on vendors and employers to be “dramatically” more transparent to allow for rigorous external audits. Absent that transparency, however, external audits must be designed around what information is publicly available. In this work we develop an external auditing methodology.

3 Methodology

3.1 Socio-technical methodology

We now present a socio-technical framework to assess the stability of algorithmic personality tests in hiring, inspired by the auditing framework of Brown *et al.* (Brown et al. 2021).

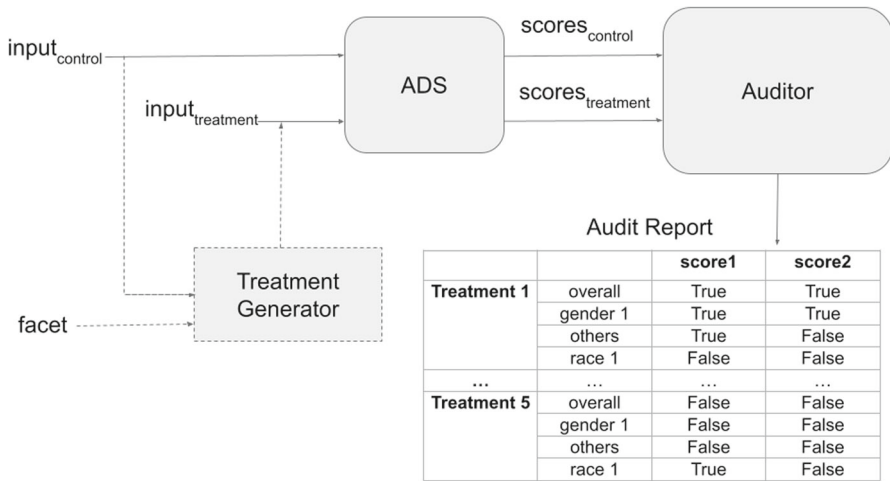


Fig. 2 Overview of the technical framework, implemented by our open-source library

1. **Define the socio-technical context** in which the system operates, and detail the system's inputs and outputs.
2. **Identify assumptions** made by the vendors regarding stability of the system.
3. **Identify key facets** of measurement across which the system assumes its outputs to be stable, based on validity assumptions.
4. **Collect data** that is representative of the tool's intended context of use.
5. **Generate treatments** by perturbing the input (control) across the features that correspond to each facet of measurement, while keeping all other features fixed to the extent possible.
6. **Identify stability metrics and acceptance/rejection criteria** that suitably capture the statistical relationship between the control and treatments.
7. **Query the external system of interest** to collect scores for the control and treated inputs.
8. **Quantify the instability across each facet** based on the selected statistical criteria.

3.2 Technical framework and open-source library

As part of this work, we developed an open-source software library³ that implements the technical components of the audit, and can be used for the stability audit of any automated decision system (ADS), given suitable input data, treatment generation techniques, and choice of stability metrics. The technical framework is shown in Fig. 2 and consists of three modules: ADS, Treatment Generator, and Auditor, described next.

³ <https://github.com/DataResponsibly/hiring-stability-audit>.

3.2.1 ADS

The ADS class is an abstraction of the algorithmic system being studied. It has a generic *score function* that takes inputs and returns scores. We include the ADS class in the audit framework to make explicit the nature of access that the user has to the system being studied, and it is intentionally designed to be generic to allow the user to model different auditing conditions. For example, for external audits that treat the system as a black box, the score function of the ADS object will be a simple look-up of the outputs that were produced by the system. As another example, external audits that do not have direct access to the system may instead fit a model on the collected data. Then, the score function will be executed over that fitted model. The framework can similarly be applied for internal audits, where the user has access to the model and invokes the score function directly.

Users can also implement custom score functions to generate *treatment baselines*. We would expect any algorithmic system that is used in the real world to be at least as good as a random guesser, and so a score function that appropriately implements random guessing can be used as a baseline for model stability, where the variation between control and treatment scores for any of the treatments should not exceed the variation observed between the control and the randomly generated outputs.

3.2.2 Treatment generator

In this work, we focus on an important desideratum of algorithmic systems—stability. We measure stability based on how robust system outputs are to different *treatments* performed on the input. The treatment generator is a technical instantiation of the mechanism that generates treatments, based on a particular *facet of measurement* of stability. In the audits implemented in this study, described in Sect. 4, the treatments are hand-designed based on domain expertise with personality scores and their use in hiring. However, we envision that future technical frameworks can at least partially automate the creation of treatments, for example, by sampling values for a particular feature from an appropriate distribution, or by automatically perturbing values in text features.

3.2.3 Auditor

The Auditor is the main class of this technical framework: it instantiates a generic Auditor that can be used to analyze the stability of a variety of algorithmic systems. The generic Auditor class allows the user to specify the following information:

1. **Score names** The framework is flexible to test the stability of multiple scores produced by the same system, and takes in a list of score names from the user. For example, *Crystal* produces four scores corresponding to the personality traits of *Dominance*, *Influence*, *Steadiness*, and *Conscientiousness*.
2. **Control score** This is the score corresponding to the unmodified/unperturbed input.

3. **Treatment scores** The framework is flexible to simultaneously analyze the stability of the ADS for several different treatments, and accepts a dictionary of treatment names and the corresponding treatment scores from the user. For example, in the audit of *Crystal*, treatments include modifying input type, modifying source context, and embedding LinkedIn URL.
4. **Demographic information.** The framework can break down audit results by demographic group. To invoke this functionality, the user can pass identifiers that link demographic information to each input and score, and specify *groups of interest* within the population being evaluated/scored by the ADS. All the subsequent analysis is then performed for both the overall population and for the specified groups.

The Auditor class currently supports the following functionality:

Statistical hypothesis testing From a socio-technical standpoint, this audit framework identifies facets across which the system assumes its output to be stable, and then tests the validity of those claims. The socio-technical heavy-lifting is in identifying these facets, designing treatments that vary along this facet, and identifying measures that capture the variation between control and treatments as a measure of instability. The Auditor class of the technical framework automates the subsequent hypothesis testing by instantiating a generic *compute_statistic()* method. This method supports several popular correlation tests (such as Spearman, Pearsons and Kendall–Tau), parametric tests (such as student-t, paired student-t and ANOVA) and non-parametric tests (such as Mann-Whitney, Wilcoxon and Kruskal–Wallis). Users can also choose to plug in their own custom functions to compute statistics that the framework does not currently support.

Users can thereby test their hypotheses about the stability of the ADS with respect to different treatments, and for demographic groups of interest, by using the Auditor class with the relevant statistical measure.

Measuring total variation Our audit framework supports functionality to compute and visualize the total variation between the control and treatments, with a large amount of variation indicating greater instability. The generic *compute_total_variation* method of the Auditor class implements this. For the purposes of the *Crystal* and *Humanitic AI* audits, we chose to measure total variation as the L1 distance between the control and each treatment, but the framework is flexible to accommodate different measures of total variation.

The Auditor class also has a *visualize_total_variation* method that produces box plots of the total variation for each treatment, broken down by demographic groups of interest. Extensions of this framework could include additional measures and visualization techniques to analyze the total variation.

Visualization The Auditor class implements the *visualize_scores* method that produces scatter plots of control vs. treatment scores, compared with the ideal $Y = X$ line (no variation between control and treatment scores), and the *visualize_total_variation* method discussed above.

4 Instantiation of the framework for personality testing in hiring

We now instantiate our socio-technical framework through external audits of *Humantic AI* and *Crystal*. Jupyter notebooks demonstrating the use of our open-source library to conduct these audits are available on GitHub, where we also publish a third audit on a synthetically-generated dataset to demonstrate a broader, more general application of the technical framework.⁴

4.1 Socio-technical context

Employers purchase candidate-screening tools from *Crystal* and *Humantic AI* and use them to build personality profiles of potential employees. Both systems offer functionality for ranking candidates based on their personality profiles. *Crystal* assigns a “job fit” score to candidates, which is measured based on a comparison to either a “benchmark candidate” with a user-specified ideal personality profile, or to a job description that is analyzed to “detect the most important personality traits.” Similarly, *Humantic AI* assigns a “match score” to candidates by comparing them to an “ideal candidate,” specified with a LinkedIn URL or an ideal personality score vector.

The hiring processes supported by these systems are not fully automated. Human decision-makers must choose whether and how to define an ideal candidate, at what stage of hiring to use the tool, and how to incorporate tool outputs into hiring decisions. For example, an HR professional may decide to use an existing employee to define an ideal candidate, then run all resumes they receive through the tool, and finally offer interviews to all candidates with match scores above 90%. A different HR department may use the system to filter resumes before human review, choosing to rank candidates based on predicted *Steadiness* scores, and then discard all but the top 25 candidates. As these examples illustrate, the human-in-the-loop implementation details are crucial to actual outcomes.

Inputs and outputs Both systems output candidate DiSC scores: vectors of 4 numeric values, each corresponding to a personality trait. *Humantic AI* produces a score for each trait on a scale from 0 to 10, while *Crystal* represents each trait as a percent of the whole, giving each a score from 0 to 100 such that all four traits sum to 100%. In addition to DiSC, *Humantic AI* also outputs scores for The Big Five model of personality.

DiSC is a behavioral psychology test that assesses the extent to which a person exhibits four personality traits: *Dominance* (D), *Influence* (I), *Steadiness* (S), and *Conscientiousness* (C).⁵ Although official DiSC documentation states that C represents *Conscientiousness*, *Humantic AI* states that C in DiSC stands for *Calculativeness*.⁶ Notably, although both *Humantic AI* and *Crystal* market DiSC as a rigorous psychology-based analysis methodology, scholarly work on DiSC in I–O psychology

⁴ <https://github.com/DataResponsibly/hiring-stability-audit>.

⁵ <https://www.discprofile.com/what-is-disc/how-disc-works>.

⁶ *Humantic AI* separately produces predictions on *Conscientiousness* within the Big Five model of personality. We posit that *Humantic AI* may have made the choice to rename the DiSC *Conscientiousness* trait to *Calculativeness* in order to avoid conflation with the Big Five trait by the same name.

has been limited, especially with regard to its validity and reliability for hiring. In fact, the DiSC website explicitly states that DiSC scores are “not recommended for pre-employment screening.”⁷

The Big Five model contains five traits: *Openness* (O), *Conscientiousness* (C), *Extraversion* (E), *Agreeableness* (A), and *Neuroticism* (N). *Humantic AI* replaces *Neuroticism* with the more palatable *Emotional Stability*, which, they explain, is “the same as *Neuroticism* rated on a reverse scale”.⁸ The use of the Big Five in personnel selection, while deemed acceptable by some I–O psychologists Goodstein and Lanyon (1999; Hurtz and Donovan 2000), is not without criticism. For example, Morgeson *et al.* argue that “the validity of personality measures as predictors of job performance is often disappointingly low” (Morgeson *et al.* 2007).

System design and validation. *Humantic AI* and *Crystal* state that they use machine learning to extract personality profiles of job candidates based on the text of their resumes and LinkedIn profiles. However, public information about model design and validation is limited. *Humantic AI* states that “all profile attributes are determined deductively and predictively from a multitude of activity patterns, metadata or other linguistic data inputs.”⁹ *Crystal* explains that their personality profiles are “predicted through machine learning and use text sample analysis and attribute analysis.”¹⁰ Neither company makes its training data publicly available or discusses the data collection and selection methodology they used. For this reason, an external audit cannot assess whether the training data is representative of the populations on which the systems are deployed.

Information about validation is limited as well. *Humantic AI* reports that their outputs “have an accuracy between 80–100%”¹¹ *Crystal* advertises that “based on comparisons to verified profiles and our user’s direct accuracy validation through ratings and endorsements, *Crystal* has an 80% accuracy rating for Predicted [sic] profiles.”¹² No additional information is given about the validation methodology, the specific accuracy metrics, or results. Finally, update schedules for the models used by the systems are not disclosed.

4.2 System assumptions

In accordance with Sloane *et al.*, our methodology is centered around testing the underlying assumptions made by algorithmic systems within their specific socio-technical context (Sloane *et al.* 2022). Because algorithmic personality tests constitute a category of psychometric instrument, they are subject to the assumptions made by the

⁷ <https://www.discprofile.com/everything-disc/hiring>.

⁸ <https://app.humantic.ai/#/candidates>.

⁹ <https://api.humantic.ai/>.

¹⁰ <https://www.crystalknows.com/blog/crystal-accuracy>.

¹¹ <https://api.humantic.ai/>.

¹² <https://www.crystalknows.com/blog/crystal-accuracy>.

traditional instruments, as laid out in Sect. 2.1. The validity of these systems is subject to the following additional assumptions:¹³

A1: The output of an algorithmic personality test is stable across input types (such as PDF or Docx) and other job-irrelevant variations in the input. This assumption corresponds to parallel forms reliability from psychometric testing (see Sect. 2.1).

A2: The output of an algorithmic personality test is stable across input sources (such as resume or LinkedIn) that are treated as interchangeable by the vendor. This assumption corresponds to cross-situational consistency (see Sect. 2.1).

A3: The output of an algorithmic personality test on the same input is stable over time. This assumption corresponds to test-retest reliability (see Sect. 2.1).

Importantly, all these assumptions are testable via an external audit. Thus, these are the assumptions on which we focus our analysis, and with respect to which we quantify stability as a necessary condition for validity.

4.3 Key facets of measurement

We identify the following key facets across which *Humantic AI* and *Crystal* operationalize reliability, as discussed in Sect. 3:

Resume file format Absent specific formatting instructions, the file format of an applicant's resume (e.g., PDF or text), should have no impact on their personality score. Per assumption **A1**, stability estimates across this facet quantify parallel forms reliability.

Source context Both systems use implicit signals within certain contexts (i.e., resumes, LinkedIn profiles, and tweets) to assign personality scores to job seekers. Further, both systems allow direct comparisons of personality scores derived from multiple source contexts, for example by ranking candidates on their "match score," which is computed from resumes for some job seekers and from LinkedIn profiles for other job seekers. Per assumption **A2**, stability estimates across this facet quantify cross-situational consistency.

Inclusion of LinkedIn URL in a resume The decision to embed a LinkedIn URL into one's resume should have no impact on the personality score computed from that resume. This is because output is expected to be stable across input sources per assumption **A2**, and across job-irrelevant input variations per **A1**.

Algorithm-time (time when input is scored). Both systems generate personality scores for the same input at different points in time, and they compare and rank job seekers based on their scores made at different times. For example, consider an extended hiring process that takes place over the course of months, with new candidates being screened at different times. In this situation, *Humantic AI* and *Crystal* would both encourage users to compare output generated months apart. Based on assumption **A3** (test-retest reliability), we expect the personality score computed on *the same input* to be the same, irrespective of when it is computed.

Participant-time (time when input is produced). An employer may keep candidate resumes on file to consider them for future positions. An HR specialist might be

¹³ Note that this list of assumptions is not exhaustive.

Table 1 Resume versions used as input

Version	File Format	Pre-Processing
Original	Various	None
De-Identified	PDF	Remove identifiers (name, phone, email, social media links, usernames). Save as PDF.
Raw Text	Raw Text	Copy text.
PDF	PDF	Save as PDF (if original in other format).
DOCX	DOCX	Remove identifiers (name, phone, email, social media links, usernames). Save as DOCX.
URL-Embedded	PDF	Remove identifiers (name, phone, email, social media accounts, LinkedIn URL). Insert hyperlinked LinkedIn URL into beginning of document. Save as PDF.

tempted to generate scores from resumes they have on file, and compare them to scores of new candidates. Neither *Humantic AI* nor *Crystal* offer any guidance to users regarding the time period during which results remain valid, thus encouraging users to generalize across participant-time. Based on **A3** (test-retest reliability), we expect the personality score computed based on time-varying input from *the same individual* to be the same, irrespective of when the input is generated.

4.4 Data collection

Primary data collection. We conducted an IRB-approved human subjects research study at New York University to seed the input corpus for the audit. For this, we recruited current graduate students at New York University's Center for Data Science ($N = 33$), Tandon School of Engineering ($N = 51$), and Courant Institute of Mathematical Sciences ($N = 10$). We further required that participants not be currently located in the European Union or the United Arab Emirates. Participants were asked to complete a survey to upload their resume, provide a link to their public LinkedIn URL, their public Twitter handle, and their demographic information. All survey questions were optional.

In total, 94 participants qualified for the study, of whom 92 submitted LinkedIn URLs, 89 submitted resumes (in PDF, Microsoft Docx, or .txt format), and 32 submitted public Twitter handles. Participants were given access to their personality profiles computed by *Crystal* and *Humantic AI* in exchange for their participation in the study. See Appendix A.1 for demographic details.

Persistent linkage of email addresses to LinkedIn profiles, and the need for de-identification. During the initial processing of participant information in *Humantic AI*, we observed that the personality profile produced from LinkedIn is often identical to the one produced from a resume containing an embedded LinkedIn URL. We hypothesized that for such URL-embedded resumes, *Humantic AI* was disregarding any information on the resume itself and pulling information from LinkedIn to generate a personality score. We further hypothesized that the system may create persistent linkages between email addresses and LinkedIn profiles.

To investigate this trend, resumes containing a LinkedIn URL and an email address were passed to *Humantic AI*. Next, we created and submitted synthetic PDF resumes, which were blank except for the email addresses that had been passed along with LinkedIn URLs, and compared the *Humantic AI* output produced by these two treatments. (Note: Due to privacy concerns, all linkage experiments used researchers' own accounts and either their own or synthetic email addresses.) It was revealed that, when *Humantic AI* encounters a document that contains both a LinkedIn URL and an email address, it persistently associates the two such that the system produces the same personality score whenever it encounters that email address in the future. Because *Humantic AI* uses the embedded URLs to import information directly from LinkedIn, the predicted profiles in our linkage experiments displayed names, photos, and employment information present on LinkedIn, but not on the resumes.

These findings further substantiate that *Humantic AI* operationalizes assumption A2 of cross-situational consistency (see Sect. 3).

These findings necessitated the use of de-identified resumes in all future *Humantic AI* experiments. De-identification allows comparison of the algorithm's predictions on resumes, without the obfuscating effect of information being pulled from LinkedIn. It also prevents participants' emails from being linked to synthetically altered versions of their resumes. See Table 1 for de-identification details. Note that de-identification was not necessary in *Crystal*, as no such linkage was observed there. Further findings from our linkage explorations are detailed in Sect. 5.1.

4.5 Treatment generation

To assess stability with respect to a facet of measurement, we need to perturb the input across the features that correspond to each facet, while keeping all other features fixed to the extent possible. As a result, we generate a pair of datasets, which we call *treatments*, for each facet. To isolate facet effects as cleanly as possible, we prepared several resume versions, described in Table 1. Details of each set of score-generating model calls that use these resume versions, or social media links, are presented in Appendix A.2. We will explain how these versions are used as treatments in the stability experiments in Sect. 5.

4.6 Stability measures

In the context of personality prediction, we identify the following measures of stability, summarized here, with additional details given in Appendix A.3.

Rank-order stability. As explained in Sect. 2.1, the reliability of psychometric instruments is measured with correlations. Thus, we select correlation as the statistical measure of rank-order stability. Morrow and Jackson make a convincing argument against providing significance levels for reliability correlations. Instead, we use the "bare minimum" of 0.90 and the "desirable standard" of 0.95, as proposed by Nunnally and Bernstein (1994), as the accept/reject threshold on correlations (Morrow et al. 1993).

Locational stability. If a system allows users to compare output across a key facet, then we should also assess locational stability across that facet, i.e., whether one facet treatment generally yields higher overall scores. We select the Wilcoxon signed-rank test, a non-parametric alternative which tests whether the median of the paired differences is significantly different than zero, as the statistical measure of locational stability. We select a suitable significance threshold after correcting for multiple hypothesis testing:

- **Bonferroni correction** controls the family-wise error rate. It is guaranteed to falsely reject the null hypothesis no more often than the nominal significance level, however, it can be overly conservative, especially when sample sizes are low (i.e., it can falsely accept the null hypothesis more often than the nominal significance level implies), refer to VanderWeele and Mathur (2019) for details:

$$\alpha_{\text{Bonferroni}} = \frac{\alpha_{\text{nominal}}}{\# \text{ tests performed}}$$

- **Benjamini–Hochberg correction** is a less conservative approach that controls the false discovery rate. The procedure ranks obtained p-values in ascending order and uses these ranks to derive corrected thresholds, which range between $\alpha_{\text{Bonferroni}}$ and α_{nominal} , refer to Benjamini and Hochberg (1995) for details:

$$\alpha_{\text{Benjamini-Hochberg}} = \frac{\text{p-value rank}}{\# \text{ tests performed}} \alpha_{\text{nominal}}$$

Total change We also identify total change as a relevant measure of instability, and use the L1 distance to measure it.

Note that these are three different ways to quantify stability, and that a system may, for example, be found to have sufficient rank-order stability but to lack locational stability, and vice versa.

4.7 Generating outputs

To conduct this audit, we purchased nine months of *Humantic AI* basic organizational membership at a total cost of \$2,250, and a combination of monthly and annual *Crystal* memberships at a total cost of \$753.82. We carried out our experiments over the period of November 23, 2020 through September 16, 2021.

One week into our evaluation, representatives from *Humantic AI* ascertained that we were using their tool to conduct an audit, and reached out to inform us that they would like to collaborate in the effort. In light of this development, we weighed the advantages and disadvantages of engaging with *Humantic AI* and decided to continue with a neutral external audit, to minimize the potential for conflicts of interest and maximize our ability to critically analyze the system for stability. The cost of that decision is that we had to forgo potential access to the underlying data, modeling decisions, features, and model parameters that a collaboration with *Humantic AI* may have afforded (Koshiyama et al. 2021; Sloane et al. 2022). While we do not have any reason to believe that the discovery of our audit caused *Humantic AI* to change their models or operation, we cannot rule out this possibility.

4.8 Computing stability measures

For the audits of *Crystal* and *Humantic AI* we compute the following statistical measures using our technical framework:

Rank Order Stability We compute Spearman's correlation (a measure of rank order stability) as follows:

```
# Instantiate the Auditor class
stability_audit = Auditor(control_scores, treatment_scores)

# Call the generic compute statistic method with parameter test set to
Spearman corr_ = stability_audit.compute_statistic(test=spearman)
["correlations"]

# Threshold correlations on desired cut-offs
corr_threshold = 0.9
corr_ > corr_threshold
```

Locational Stability. We perform Wilcoxon's signed rank test (as the measure of locational stability) as follows:

```
# This time calling compute statistic method with parameter test set to
Wilcoxon

pvals = stability_audit.compute_statistic(test=wilcoxon)
["p_values"]

# Using an alpha of 0.05
alpha_threshold = 0.05

# Correct for multiple hypothesis testing using Benjamini-Hochberg correction
corrected = stability_audit.multiple_hypothesis_correction(
    pvals, alpha = alpha_threshold,
    method='fdr_bh')

# Threshold pvalues on desired cut-off
corrected > alpha_threshold
```

Total Variation. We also compute the L1 distance (as a measure of total variation) as follows:

```
# This time calling the compute total variation method, whose default measure
is the L1 norm

total_variation = stability_audit.compute_total_variation()
```

Table 2 Summary of stability results for *Crystal* and *Humantic AI*, with respect to facets of measurement from Sect. 4.3.

Facet	<i>Crystal</i>	<i>Humantic AI</i>	Details
Resume file format	×	✓	Sect. 5.3
LinkedIn URL in resume	?	×	Sect. 5.4
Source context	×	×	Sect. 5.5
Algorithm-time / immediate	✓	✓	Sect. 5.6
Algorithm-time / 31 days	✓	×	Sect. 5.6
Participant-time / LinkedIn	×	×	Sect. 5.7
Participant-time / Twitter	N/A	✓	Sect. 5.7

“✓” indicates both sufficient rank-order stability ($r \geq 0.90$) and sufficient locational stability ($p \geq \alpha_{\text{Benjamini-Hochberg}}$) in all traits, “✗” indicates either insufficient rank-order stability ($r < 0.90$) or significant locational instability ($p < \alpha_{\text{Benjamini-Hochberg}}$) in at least one trait, and “?” indicates the facet was not tested in our audit

5 Results

Table 2 summarizes the results of our audit. We found that *Humantic AI* and *Crystal* predictions both exhibit rank-order instability with respect to source context and participant-time. In addition, *Crystal* is rank-order unstable with respect to file format, and *Humantic AI* is rank-order unstable with respect to URL-embedding in resumes. The systems were sufficiently rank-order stable with respect to all other facets. We did not find any significant locational instability in *Crystal*. Some traits in *Humantic AI* displayed significant locational instability with respect to URL-embedding, source context, and participant-time. Complete experimental results can be found in Appendix B.

5.1 Persistent linkage and privacy violations in *Humantic AI*

Investigative linkage experiments revealed that when *Humantic AI* encounters a document that contains a LinkedIn URL and an email address, the resulting profile will have a 100% confidence score, and it will contain information found only on LinkedIn (including name, profile picture, and job descriptions and dates). Furthermore, the *Humantic AI* model produces the same personality profile whenever it encounters that email address in the future. This linkage persists regardless of how different the new resume is from the one that initially formed the linkage. The email address in question need not be associated with the LinkedIn profile, or even with the candidate. We observed one case in which a participant listed contact information for references, and *Humantic AI* created a link between a reference’s email and the participant’s LinkedIn.

We also found that, once a linkage between an email address and a LinkedIn URL had been made, we were able to alter the personality score produced from a LinkedIn profile by submitting a resume with strong language, namely, containing keywords “sneaky” and “adversarial.” We therefore conclude that the linkage is used by *Humantic AI* in both directions: the content of a LinkedIn profile can affect the personality

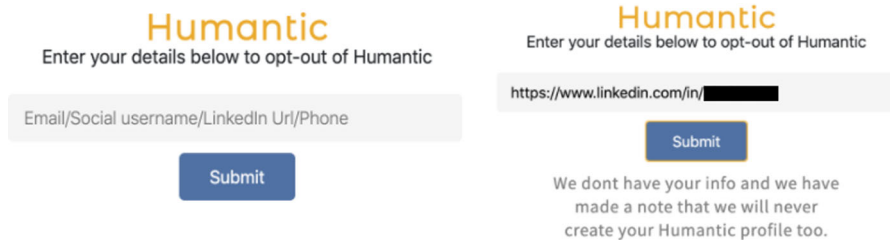


Fig. 3 Screen shots of the *Humantic AI* “opt out” feature

score computed from a linked resume, and the content of a linked resume can affect personality score computed based on a LinkedIn profile.

We did not observe any linkage with participants’ Twitter accounts. However, when we used high-profile celebrity Twitter accounts as input, *Humantic AI* produced profiles that contained links to several other profiles, including Google+, LinkedIn, Facebook, and Klout. We observed one case in which a high-profile popstar was linked to a software engineer of the same name.

Although *Humantic AI* offers an option at the bottom of their website to “opt out of *Humantic AI*” by entering an email, social network username, LinkedIn URL, or phone number (see Fig. 3), this feature seems to be inoperable. Various forms of participant information were entered into this field, yet, personality scores associated with this information in the past persisted on the *Humantic AI* dashboard, and new results were returned when the information was passed to *Humantic AI* in a new account. In cases where LinkedIn profiles were deactivated after profiles were created from them, it was observed that *Humantic AI* would still create new profiles from the deactivated LinkedIn, even on different *Humantic AI* accounts.

5.2 Score distributions

Output scores in *Humantic AI* were approximately normally distributed, with the exception of DiSC *Calculativeness*, which was strongly left-skewed in all runs.

We observed discontinuity in *Crystal* output, which was particularly marked in *Steadiness* and *Conscientiousness*, as shown in Fig. 4. For example, no one in our sample had a *Steadiness* score between 40–50, but many individuals had scores in the 20–30 range, and then again in the 55–65 range. This may be problematic from the point of view of stability, because a small change in the input may lead to a large change in output across the point of discontinuity, effectively moving between clusters. In fact, we observe this in Fig. 4, where in two cases, the value of *Steadiness* jumps from around 30 for raw text resumes to around 60 for PDF resumes. Having a PDF resume can make you twice as steady, according to *Crystal*. Yet, two other examples show the opposite effect: A raw text resume scores about twice as high on *Steadiness* compared to PDF, for another pair of individuals in our sample. And so having a PDF resume can also make you half as steady, according to *Crystal*. There are further examples of this for *Conscientiousness*, also shown in Fig. 4.

We found no evidence of significant locational instability in *Crystal*. The median for each DiSC trait remained fairly constant across all *Crystal* runs. The median

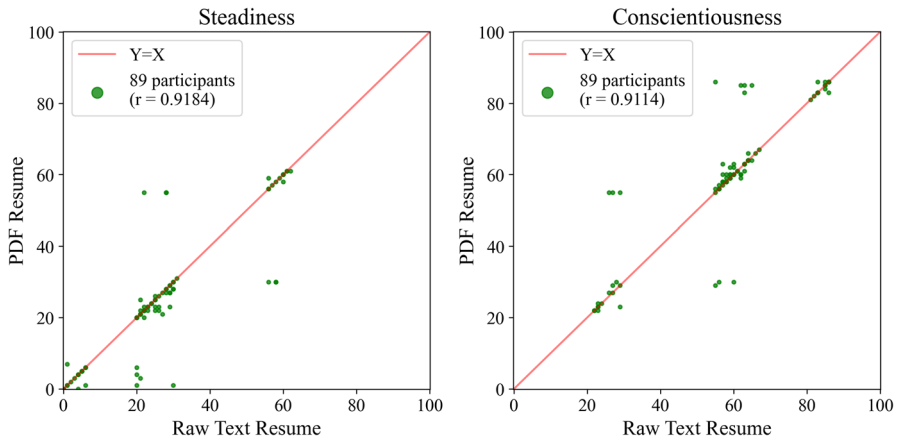


Fig. 4 Comparison of *Crystal* output across the resume file format facet. Note evidence of discontinuous measurement in DiSC *Steadiness* and *Conscientiousness*, with some participants' scores moving between clusters with different file formats

Dominance score was always 5, the median *Influence* score was always 10, the median *Steadiness* score was always 22 or 23, and the median *Conscientiousness* score ranged from 59 to 62.

5.3 File format

We determine that *Humantic AI* is in general sufficiently stable with respect to file format. Rank correlations range from 0.982 (*Emotional Stability*) to 0.998 (*Steadiness*). (The two sets of runs are constant with regard to participant-time, and are very close to each other in terms of algorithm-time; scores for the de-identified PDF and Docx resumes were generated on the same day, within minutes of each other.)

Crystal's overall stability across the file format facet fails to meet Nunnally and Bernstein's preferred standard of 0.95 for *Steadiness* (0.918) and *Conscientiousness* (0.911), and falls below the minimum limit of 0.90 for *Dominance* (0.822) and *Influence* (0.826). In some subgroups, *Steadiness* and *Conscientiousness* do fall below 0.90: female ($N = 33$) and those whose primary language is English ($N = 56$). Although PDF resumes were scored by *Crystal* four months earlier than raw text resumes, given the perfect reproducibility of *Crystal's* text predictions, albeit over a shorter time span, we can assume that algorithm-time is not a factor here.

There were no significant locational stability differences across the file format facet in either *Humantic AI* or *Crystal*.

5.4 Inclusion of LinkedIn URL in resume

We discovered substantial instability with regard to URL-embedding in resumes in *Humantic AI*. Correlations between de-identified resumes and the same resumes with LinkedIn URLs embedded into them ranged from 0.077 (*Extraversion*) to 0.688

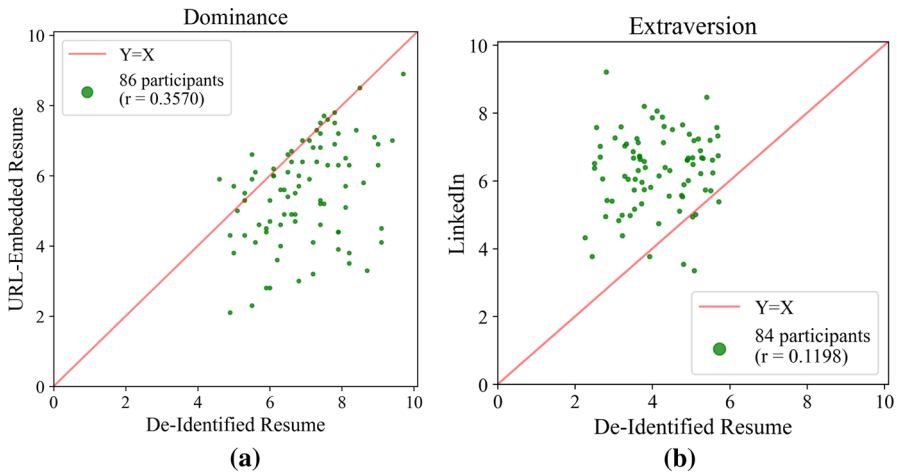


Fig. 5 **a** *Humantic AI Dominance* scores from de-identified and URL-embedded resumes. **b** *Humantic AI Extraversion* scores produced by de-identified resumes and LinkedIn profiles

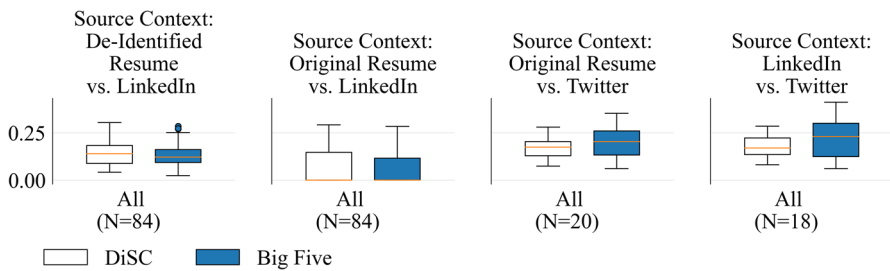


Fig. 6 Normalized L1 distances between *Humantic AI* DiSC and Big Five scores produced from pairs of treatments that vary with respect to their input source

(*Calculativeness*). We also discovered locational differences deemed significant by the Bonferroni threshold in *Dominance*, *Steadiness*, Big Five *Conscientiousness*, *Extraversion*, and *Agreeableness*. Under the more liberal Benjamini–Hochberg standard, there were also significant locational differences in DiSC *Calculativeness* and *Openness*. Figure 5a gives a representative example; complete results are presented in Appendix B.1.

We note that algorithm-time is unfortunately an unavoidable factor here; the two resume versions were run about four months apart. Furthermore, if we accept that *Humantic AI* uses information from LinkedIn profiles when it encounters embedded LinkedIn URLs, then we are also faced with a mismatch in participant-time.

5.5 Source context

Humantic AI and *Crystal* both displayed low stability across input sources. See Figure 6 for comparison of L1 distances between each treatment of the input source facet in *Humantic AI*.

Crystal's rank-order correlations between PDF resumes and LinkedIn profiles were all below the 0.90 threshold; they ranged from 0.233 (*Dominance*) to 0.526 (*Influence*). There was no significant locational instability in *Crystal*. PDF resumes and LinkedIn URLs were scored the same day, and, as we will discuss in Sect. 5.6, *Crystal* is immediately reproducible, and so we can rule out algorithm-time as a factor in this finding. Furthermore, for each candidate, this scoring took place within two weeks of resumes being submitted; thus, the participant-time of the resume matches very nearly to the participant-time of the LinkedIn. With all other facets being identical or near-identical, we can safely attribute the observed score differences to differences in source context.

De-identified resumes were submitted to *Humantic AI* 4 months after LinkedIn profiles had been run. This difference in algorithm-time hampers our interpretation of cross-profile correlations. Nonetheless, it is undeniably troublesome that the observed correlations are as low as 0.090 (*Dominance*), and that there were significant locational differences under Bonferroni in *Dominance* and *Extraversion*, and under Benjamini–Hochberg in *Steadiness* and *Openness*. See Appendix B.2 for details.

We can avoid the issue of algorithm-time by using *Humantic AI* scores derived from original resumes, which were run at the same time as LinkedIn profiles. However, these results are somewhat misleading, as 57 of the 84 resumes in this experiment contained some form of LinkedIn URL. Considering the evidence that *Humantic AI* uses information directly from LinkedIn in such cases, correlations derived from original resumes are likely to overestimate cross-contextual stability. Nevertheless, the correlations we observe across all 84 participants range from 0.177 (*Dominance*) to 0.712 (Big Five *Conscientiousness*), with significant locational differences under Bonferroni in *Dominance* and *Extraversion*; and in *Influence* and Big Five *Conscientiousness* under Benjamini–Hochberg. We also found significant differences for non-native English speakers in *Agreeableness* under Benjamini–Hochberg. See Appendix B.2 for details. Limiting analysis to the 27 participants whose original resumes contained no reference to LinkedIn, we find that the correlations straddle zero, ranging from -0.310 (*Influence*) to 0.297 (DiSC *Calculativeness*).

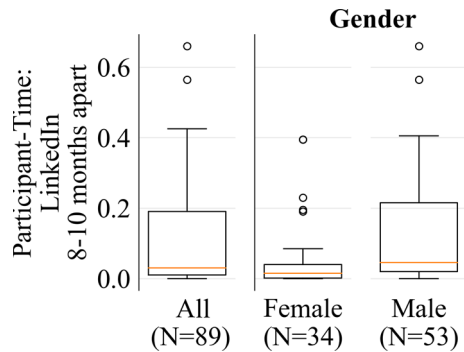
Figure 5b highlights some of these results. Appendix B.2 presents details of this experiment, and further includes a comparison of *Humantic AI* scores computed from Twitter to those computed from original resumes and from LinkedIn.

5.6 Algorithm-time

Crystal results on resumes were reproducible immediately as well as one month later. We can conclude that *Crystal*'s text prediction tool is deterministic and was not updated over the course of April 2021, when the experiment was performed.

Humantic AI results were not perfectly reproducible, even immediately. This may be explained by a non-deterministic prediction function, or by an online model that is updated with each prediction it makes. The latter explanation is in-line with our findings in the linkage investigations, where we observed that one call to the model can influence the outcome of other calls. Only *Steadiness* and DiSC *Calculativeness* remained constant for all participants when identical resumes were run back-to-back.

Fig. 7 Normalized L1 distances between *Crystal* DiSC scores produced from LinkedIn profiles scored 8–10 months apart



One participant had changes in their *Dominance* and *Influence* scores (DiSC total normalized L1 difference was 0.005), and two participants had changes in their Big Five scores (maximum Big Five total normalized L1 difference was 0.003). The correlations for immediate reproducibility were all above 0.95, and there were no significant locational differences.

After 31 days, rank-order correlations in *Humantic AI* ranged from 0.962 (*Extraversion*) to 0.998 (*DiSC Calculativeness*). Although the overall *Humantic AI* correlations across algorithm-time were all above the 0.95 threshold, we find that for non-native English speakers ($N = 33$), *Dominance* ($r = 0.946$) and *Extraversion* ($r = 0.934$) both fell below 0.95. We also find significant instability in *Openness* under Benjamini–Hochberg.

See Appendix B.3 for additional details about this experiment.

5.7 Participant-time

Humantic AI scores on Twitter accounts showed no change over 7–9 months. LinkedIn correlations across 7–9 months of participant-time were all below the 0.90 threshold: they ranged from 0.225 (*Dominance*) to 0.768 (*Emotional Stability*). Under Bonferroni correction, we found a significant difference in Big Five *Conscientiousness* scores, and under Benjamini–Hochberg we found a significant difference in *Agreeableness*.

Crystal LinkedIn correlations across 8–10 months of participant-time were all below the 0.90 threshold as well, ranging from 0.531 (*Dominance*) to 0.868 (*Steadiness*). We found that the reliability for male participants was particularly low ($N = 53$, $r = 0.232$). See Figure 7 for cross-gender comparison of L1 distances between participant-time treatments. There was no significant locational instability across participant-time in *Crystal*. See Appendix B.4 for additional details about this experiment.

6 Discussion

6.1 Stability audit conclusions

Humantic AI and *Crystal* both exhibit low reliabilities across time and input source context. *Humantic AI* also exhibited low reliability with respect to the presence of LinkedIn URLs in resumes. *Crystal*'s reliability with respect to resume format is unacceptably low as well. The correlations we observed allow us to conclude that the tools cannot be considered valid instruments in high-stakes decisions.

Overall, each of these observed unreliabilities undermines the cost and effort reduction that employers seek from candidate screening tools. Employers' desire for valid decisions reflective of job performance is severely compromised by sensitivity to job-irrelevant factors. Thus, we find that *Humantic AI*'s sensitivities to participant-time, URL-embedding, and source context, and *Crystal*'s sensitivities to file format and source context, could be quite problematic for employers. The sensitivity of these algorithms to job-irrelevant factors is also a threat to individual fairness; a job seeker could reasonably conclude from the present audit that *Humantic AI* and *Crystal* are both likely to judge their job-worthiness unfairly, letting meaningless criteria dictate their outcomes.

These unreliabilities are also at odds with the trustworthiness that society seeks in its AI products. *Humantic AI*'s lack of reproducibility is a particularly insidious violation of trustworthiness, because it undermines the power of audits on its system. Although *Humantic AI*'s stability over algorithm-time exceeds Nunnally and Bernstein's classical 0.95 reliability threshold for tests used to make decisions about individuals (see Sect. 2.1), the Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK have asserted that reproducibility is "a mandatory condition for reliability" (Riksrevisjonen 2020). Irreproducibility resulting from frequently or continuously updated models poses a threat to the ongoing monitoring and auditing necessary to ensure a system is working as expected (Bogen and Rieke 2018; Koshiyama et al. 2021).

Finally, *Humantic AI*'s and *Crystal*'s lack of transparency regarding training data and model architecture are at odds with privacy concerns. *Humantic AI*'s deceptive and ineffective opt-out option is an example of what Ajunwa calls "algorithmic blackballing," whereby an applicant's profile is allowed to live on past its shelf-life (Ajunwa 2021). This is especially dangerous in combination with the potential to leverage *Humantic AI*'s email linkage mechanism in an adversarial attack. *Humantic AI*'s failed opt-out option may also violate the California Consumer Privacy Act's right to delete (California Civil Code 2018).

6.2 Study limitations

In our audit we do not conduct stakeholder evaluations. Several audits and frameworks emphasize the importance of stakeholder evaluation and impact assessment (Brown et al. 2021; Fjeld et al. 2020; ORCAA 2020; Raji et al. 2020; Razavi et al. 2021; Sloane et al. 2022).

For example, Metcalf *et al.* explain that an external audit must not stand in as an algorithmic impact assessment (Metcalf et al. 2021). Without collaboration of internal agents, third parties do not have access to design decisions or stakeholder interviews, and cannot directly influence change in the design or operation of the algorithm should it be needed. Per Ajunwa, algorithms need to be audited internally as well as externally (Ajunwa 2021).

Although this audit considers various dimensions of reliability and stability, the analysis is not comprehensive. We have constrained our audit methodology to analyze the numerical scores produced by personality prediction AI that claim to offer a quantitative measure of personality, such as the DiSC and Big Five scores produced by *Crystal* and *Humantic AI*. However, much of the advertising of such tools focus on the profiles holistically, not just on the scores. Further *Crystal* and *Humantic AI* both categorize candidates into one of several types and produce descriptive personality profiles. Written profiles are likely influential in hiring decisions, however, in the interest of keeping the scope of our work manageable, we leave a treatment of stability in these textual profiles to future work.

The audit methodology is also limited by its emphasis on comparing pairs of control and treatment scores. For example: *Humantic AI* often fails to produce profiles from inputs (see the discrepancies between number of inputs submitted and number of profiles produced in Table 5). This is especially common when using Twitter profiles. By simply disregarding the failed inputs, we may be introducing some sampling bias into our results. Furthermore, such non-results may exhibit problematic biases (ORCAA 2020).

Our study population was constrained to technical graduate students at NYU, studying in the realms of computer and data science. This was done in an attempt to control for differences in algorithm response due to characteristics such as job field, experience level, and writing style. We also felt that this restriction more closely replicated a pool of candidates who might realistically be compared to one another in a job search. However, this narrowness, and our modest cohort size ($N = 94$), restrict the generalizability of the results of our audit of *Crystal* and *Humantic AI*.

Additionally, this audit evaluates only the intermediate personality profile results, and does not relate them to hiring outcomes. Our audit did not use the “job fit” or “match score” features because, as external auditors, we did not have access to information on how ideal candidates are defined or how thresholds are set. Without this information, we cannot assess outcomes-based fairness metrics. This means that critical questions of discrimination remain out of scope for this study. We caution that the adverse impact of human-in-the-loop hiring systems must be assessed on an employer-by-employer basis in order to account for crucial implementation details and differences in the context of use.

7 Conclusions and future work

In this paper, we investigated the reliability of algorithmic personality tests used in hiring. We gave an overview of the key literature on psychometric testing applied to hiring and in algorithm auditing, and found that, although reliability is seen as a

necessary condition for the validity of a psychometric instrument, it has not received substantial treatment in algorithm audits. Based on this observation, we developed a socio-technical audit methodology, informed by psychometric theory and sociology, to test the stability of black-box algorithms that predict personality for use in hiring. We also developed an open-source software library to automate the quantitative components of this framework. We then instantiated this methodology in an external audit of two systems, *Humantic AI* and *Crystal*, using a dataset of job applicant profiles collected through an IRB-approved study. Using our audit methodology, we found that both systems lack reliability across key facets of measurement, and concluded that they cannot be considered valid personality assessment instruments.

The present study demonstrates that stability, though often overlooked in algorithm audits, is an accessible metric for external auditors. We found that stability is highly relevant to the application of personality prediction. Furthermore, because reliability is a prerequisite of validity, stability is in fact relevant whenever validity is. Importantly, we note that, while reliability is a necessary condition for validity, it is not a sufficient condition. Further evidence of domain-specific validity is essential to support the use of algorithmic personality tests in hiring.

Our methodology can be used by employers to make informed purchasing and usage decisions, and to better interpret algorithm outputs, by legislators to guide regulation, and by consumers to make informed decisions about how and when to disclose their information to potential employers. Our open-source software library reduces the amount of effort that would be required to conduct such analyses.

Moreover, given its modular design, our software library can be easily extended to support other reliability-related measures. As mentioned in Sect. 3, we envision an extension of the *Treatment Generator* that can (at least partially) automate the creation of treatments, for example, by sampling values for a particular feature from an appropriate distribution, or by automatically perturbing values in text features. The library's visualization capabilities, which currently include scatterplots and boxplots, can also be extended to facilitate the generation of audit results that are amenable to a wide variety of stakeholders, both technical and non-technical. The library already computes statistics broken down by demographic groups of interest, and can also easily be extended to compute fairness-related measures.

Algorithmic audits must not be one-size-fits-all. The tendency of auditors, especially within the hiring domain, to rely on legal frameworks as a scoping mechanism is likely to leave important risks undetected. Current legal frameworks are insufficient; furthermore, legality does not equate to ethics. Instead, we recommend that auditors interrogate the assumptions operationalized by systems, and design audits accordingly.

Finally, we note that this work was conducted by an interdisciplinary team that included computer and data scientists, a sociologist, an industrial psychologist, and an investigative journalist. This collaboration was both necessary and challenging, requiring us to reconcile our approaches and methodological toolkits, forging new methods for interdisciplinary collaboration.

Acknowledgements We thank Dhara Mungra for her work on data collection and preliminary analysis, and Daphna Harel and Joshua Loftus for their advice on statistical methods.

Funding This research is supported in part by NSF Awards No. 1934464, 1922658, and 1916505, and by Underwriters Laboratories Inc. through the Center for Advancing Safety of Machine Intelligence (PI Stoyanovich), and by the NYU Center for the Humanities Digital Humanities Seed Grant (PIs Schellmann and Sloane).

Availability of data and material Anonymized datasets generated for and analysed during the current study are available at <https://github.com/DataResponsibly/hiring-stability-audit/tree/main/data>.

Code Availability Code for data cleaning and analysis is provided for replication. It is available at <https://github.com/DataResponsibly/hiring-stability-audit>.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval All procedures performed involving human participants were in accordance with the ethical standards of the NYU institutional review board and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Consent to participate Informed consent was obtained from all individual participants included in the study.

Consent for publication Participants granted informed consent to publish information not containing identifiers, including personality profile results and demographic data.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

A Additional audit details

A.1 Participant demographics

This section supplements the description of the *primary data collection* discussed in Sect. 4.4 with information about participant demographics. A total of 94 participants were enrolled in our IRB-approved study. 88% of participants were pursuing a Master's

Table 3 Demographics of study participants: gender and race

	Gender			Race			
	Male	Female	Other	Asian	White	Other	No answer
N	56	36	2	57	24	12	1
%	59.6	38.3	2.1	60.6	25.5	12.8	1.0

Table 4 Demographics of study participants: birth country and primary language

	Birth Country					Primary Language	
	India	USA	China	Other	No answer	English	Other
N	34	28	12	18	2	60	34
%	36.2	29.8	12.8	19.1	2.1	63.8	36.2

Degree and 12% were pursuing a PhD degree. Their ages ranged from 21–40 with a mean of 26. Tables 3 and 4 provide additional details regarding participants' gender, race, birth country, and primary language.

A.2 Treatments for each facet

Details of score-generating model calls to generate treatments for each facet, discussed in Sect. 4.4, are presented in Table 5 for *Humantic AI* and in Table 6 for *Crystal*. In these tables, we list the type of input (e.g., Original Resume or LinkedIn profile), the identifier of the run that corresponds to this input, and the range of dates over which the system (*Humantic AI* or *Crystal*) was invoked on this type of input. We also

Table 5 *Humantic AI* runs (i.e., sets of score-generating calls to *Humantic AI* models)

Input type	Run ID	Run dates	# Inputs	# Outputs
Original Resume	HRo1	11/23/2020 - 01/14/2021	89	88
De-Identified Resume	HRi1	03/20/2021 - 03/28/2021	89	89
De-Identified Resume	HRi2	04/20/2021 - 04/28/2021	89	89
De-Identified Resume	HRi3	04/20/2021 - 04/28/2021	89	89
DOCX Resume	HRd1	03/20/2021 - 03/28/2021	89	89
URL-Embedded Resume	HRu1	04/09/2021 - 04/11/2021	86	86
LinkedIn	HL1	11/23/2020 - 01/14/2021	92	88
LinkedIn	HL2	08/10/2021 - 08/11/2021	92	91
Twitter	HT1	11/23/2020 - 01/14/2021	32	21
Twitter	HT2	08/10/2021 - 08/11/2021	32	21

Table 6 *Crystal* runs (i.e., sets of score-generating calls to *Crystal* models)

Input type	Run ID	Run dates	# Inputs	# Outputs
Raw Text Resume	CRr1	03/31/2021 - 04/02/2021	89	89
Raw Text Resume	CRr2	05/01/2021 - 05/03/2021	89	89
Raw Text Resume	CRr3	05/01/2021 - 05/03/2021	89	89
PDF Resume	CRp1	11/23/2020 - 01/14/2021	89	89
LinkedIn	CL1	11/23/2020 - 01/14/2021	92	91
LinkedIn	CL2	09/13/2021 - 09/16/2021	89	89

list input size (“Inputs Submitted”) and output size (“Profiles Produced”). Note that output size may be smaller compared to input size, and sometimes substantially so. For example, for runs HT1 and HT2, we used 32 Twitter handles as input to *Humantic AI*, but we took only 21 personality profiles produced as output into consideration. This is because *Humantic AI* did not produce personality profiles from the remaining 11 accounts, but instead returned errors saying the Twitter profiles were “thin.”

A.3 Choice of stability metrics

This section describes the metrics used to assess facet-specific stability.

Rank-order stability Because DiSC scores were discontinuous in *Crystal*, we use Spearman rank correlation rather than Pearson’s correlation coefficient to quantify rank-order stability. Rank-order stability results are presented in Tables 7, 8, and 9.

Locational stability Similarly, we use the Wilcoxon signed-rank test to assess the significance of paired differences. Unlike the Student’s t-test, the Wilcoxon signed-rank test does not assume the data is normally distributed. Locational stability results can be found in Tables 10, 11, and 12. We start with a nominal α of 0.05. In *Crystal*, we test the median change of the four DiSC traits across five facets, for a total of 20 tests and a Bonferroni-corrected α of 0.0025. In *Humantic AI*, we test the Big Five traits and the four DiSC traits across eleven facets, for a total of 99 tests and a Bonferroni-corrected α of 5.05×10^{-4} .

Total change To compute total change, we calculate the L1 distance between the output vectors of the two runs for each subject. In order to compare results across different scales, this distance is normalized by the total range of output space. The normalization constant is the inverse of the sum of possible score ranges for each trait in the category. For example, *Humantic AI* produces four DiSC scores

Table 7 Rank-order stability of *Crystal* DiSC scores, as measured by Spearman’s rank correlations. Columns labeled D (*Dominance*), I (*Influence*), S (*Steadiness*), C (*Conscientiousness / Calculativeness*).

Facet	Input Versions	N	D	I	S	C
File Format	Raw Text vs. PDF Resume (CRr1 vs. CRp1)	89	0.8225	0.8260	<i>0.9184</i>	<i>0.9114</i>
Source Context	PDF Resume vs. LinkedIn (CRp1 vs. CL1)	86	0.2335	0.5258	0.5103	0.3585
Immediate Rep.	Raw Text Resume back-to-back (CRr2 vs. CRr3)	89	1.0000	1.0000	1.0000	1.0000
Algorithm-Time	Raw Text Resume 31 days apart (CRr1 vs. CRr2)	89	1.0000	1.0000	1.0000	1.0000
Participant-Time	LinkedIn 8–10 months apart (CL1 vs. CL2)	89	0.5314	0.7062	0.8676	0.7811

Reliabilities below 0.90 highlighted in bold; those between 0.90 and 0.95 highlighted in italic. Results are discussed in Sects. 5.3, 5.4, 5.5, 5.6, and 5.7

Table 8 Rank-order stability of *Humantic AI* DiSC scores, as measured by Spearman’s rank correlations. Columns labeled D (*Dominance*), I (*Influence*), S (*Steadiness*), C (*Conscientiousness / Calculativeness*).

Facet	Input Versions	N	D	I	S	C
File Format	De-Identified Resume vs. DOCX Resume (HRi1 vs. HRd1)	89	0.9956	0.9924	0.9978	0.9959
URL Embedding	URL-Embedded Resume vs. De-Identified Resume (HRu1 vs. HRi1)	86	0.3570	0.6253	0.5480	0.6878
URL Embedding	URL-Embedded Resume vs. LinkedIn (HRu1 vs. HL1)	83	0.1555	0.3382	0.6074	0.4701
Source Context	De-Identified Resume vs. LinkedIn (HRi1 vs. HL1)	84	0.0903	0.2553	0.3941	0.3331
Source Context	Original Resume vs. LinkedIn (HRo1 vs. HL1)	84	0.1775	0.4016	0.6939	0.6249
Source Context	Original Resume vs. Twitter (HRo1 vs. HT1)	20	-0.5211	0.1026	0.0382	-0.1475
Source Context	LinkedIn vs. Twitter (HL1 vs. HT1)	18	-0.1317	0.0203	-0.1120	-0.4329
Immediate Rep.	De-Identified Resume back-to-back (HRi2 vs. HRi3)	89	0.9999	1.0000	1.0000	1.0000
Algorithm-Time	De-Identified Resume 31 days apart (HRi1 vs. HRi2)	89	0.9726	0.9948	0.9925	0.9980
Participant-Time	LinkedIn 7–9 months apart (HL1 vs. HL2)	88	0.2248	0.4186	0.6597	0.5827
Participant-Time	Twitter 7–9 months apart (HT1 vs. HT2)	21	1.0000	1.0000	1.0000	1.0000

Reliabilities below 0.90 highlighted in bold. Results are discussed in Sects. 5.3, 5.4, 5.5, 5.6, and 5.7

Table 9 Rank-order stability of *Humantic AI* Big Five scores, as measured by Spearman’s rank correlations. Columns labeled O (*Openness*), C (*Conscientiousness*), E (*Extraversion*), A (*Agreeableness*), and S (*Emotional Stability*).

Facet	Input Versions	N	O	C	E	A	S
File Format	De-Identified vs. DOCX Resume (HRi1 vs. HRd1)	89	0.9891	0.9936	0.9939	0.9927	0.9816
URL Embedding	URL-Embedded vs. De-Identified Resume (HRu1 vs. HRi1)	86	0.3988	0.3845	0.0772	0.4190	0.4040
URL Embedding	URL-Embedded vs. LinkedIn (HRu1 vs. HL1)	83	0.6381	0.5470	0.5786	0.6839	0.7018
Source Context	De-Identified Resume vs. LinkedIn (HRi1 vs. HL1)	84	0.2180	0.1558	0.1198	0.2020	0.2186

Table 9 continued

Facet	Input Versions	N	O	C	E	A	S
Source Context	Original Resume vs. LinkedIn (HRo1 vs. HL1)	84	0.5985	0.7124	0.5827	0.6136	0.5990
Source Context	Original Resume vs. Twitter (HRo1 vs. HT1)	20	-0.1768	0.2324	-0.1128	-0.2316	0.0692
Source Context	LinkedIn vs. Twitter (HL1 vs. HT1)	18	-0.2158	0.0000	-0.1559	-0.1517	-0.1125
Immediate Rep.	De-Identified Resume back-to-back (HRi2 vs. HRi3)	89	1.0000	1.0000	1.0000	0.9999	1.0000
Algorithm-Time	De-Identified Resume 31 days apart (HRi1 vs. HRi2)	89	0.9954	0.9969	0.9618	0.9921	0.9854
Participant-Time	LinkedIn 7–9 months apart (HL1 vs. HL2)	88	0.6879	0.6928	0.7301	0.7518	0.7678
Participant-Time	Twitter 7–9 months apart (HT1 vs. HT2)	21	1.0000	1.0000	1.0000	1.0000	1.0000

Reliabilities below 0.90 highlighted in bold. Results are discussed in Sects. 5.3, 5.4, 5.5, 5.6, and 5.7

Table 10 Significance in locational instability of *Crystal* DiSC scores, as measured by two-tailed Wilcoxon signed-rank test p-values. Columns labeled D (*Dominance*), I (*Influence*), S (*Steadiness*), C (*Conscientiousness / Calculativeness*).

Facet	Input Versions	N	D	I	S	C
File Format	Raw Text vs. PDF Resume (CRr1 vs. CRp1)	89	0.5026	0.4208	0.0173	0.0370
Source Context	PDF Resume vs. LinkedIn (CRp1 vs. CL1)	86	0.4190	0.0012	0.7010	0.8421
Immediate Rep.	Raw Text Resume back-to-back (CRr2 vs. CRr3)	89	N/A	N/A	N/A	N/A
Algorithm-Time	Raw Text Resume 31 days apart (CRr1 vs. CRr2)	89	N/A	N/A	N/A	N/A
Participant-Time	LinkedIn 8–10 months apart (CL1 vs. CL2)	89	0.7299	0.6518	0.3305	0.2870

The absence of bold highlighting indicates that all values are below both the Benjamini–Hochberg and Bonferroni-corrected thresholds based on α of 0.05. “N/A” values reflect experiments where there was zero change across the facet. Results are discussed in Sects. 5.3, 5.4, 5.5, 5.6, and 5.7

Table 11 Significance in locational instability of *Humantic AI* DiSC scores, as measured by two-tailed Wilcoxon signed-rank test p-values. Columns labeled D (*Dominance*), I (*Influence*), S (*Steadiness*), C (*Conscientiousness / Calculativeness*).

Facet	Input Versions	N	D	I	S	C
File Format	De-Identified vs. DOCX Resume (HRi1 vs. HRd1)	89	0.2510	0.2940	0.4574	0.2539
URL Embedding	URL-Embedded vs. De-Identified Resume (HRu1 vs. HRi1)	86	0.0000	0.3194	<i>0.0005</i>	<i>0.0047</i>
URL Embedding	URL-Embedded Resume vs. LinkedIn (HRu1 vs. HL1)	83	<i>0.0066</i>	0.1825	0.5324	0.1213
Source Context	De-Identified Resume vs. LinkedIn (HRi1 vs. HL1)	84	0.0000	0.0580	<i>0.0013</i>	0.3259
Source Context	Original Resume vs. LinkedIn (HRo1 vs. HL1)	84	0.0000	<i>0.0050</i>	0.2299	0.5911
Source Context	Original Resume vs. Twitter (HRo1 vs. HT1)	20	0.5706	0.3118	0.1975	0.6874
Source Context	LinkedIn vs. Twitter (HL1 vs. HT1)	18	0.0342	0.3247	0.6095	0.5539
Immediate Rep.	De-Identified Resume back-to-back (HRi2 vs. HRi3)	89	0.3173	0.3173	N/A	N/A
Algorithm-Time	De-Identified Resume 31 days apart (HRi1 vs. HRi2)	89	0.1416	0.5971	0.5690	0.0307
Participant-Time	LinkedIn 7–9 months apart (HL1 vs. HL2)	88	0.0709	0.0800	0.3457	0.2969
Participant-Time	Twitter 7–9 months apart (HT1 vs. HT2)	21	N/A	N/A	N/A	N/A

Bold highlighting indicates value below Bonferroni-corrected threshold based on α of 0.05. Italic indicates p-value below Benjamini–Hochberg corrected threshold and above Bonferroni-corrected threshold. “N/A” values reflect experiments where there was zero change across the facet. Results are discussed in Sects. 5.3, 5.4, 5.5, 5.6, and 5.7

Table 12 Significance in locational instability of *Humantic AI* Big Five scores, as measured by two-tailed Wilcoxon signed-rank test p-values. Columns labeled O (*Openness*), C (*Conscientiousness*), E (*Extraversion*), A (*Agreeableness*), and S (*Emotional Stability*).

Facet	Input Versions	N	O	C	E	A	S
File Format	De-Identified vs. DOCX Resume (HRi1 vs. HRd1)	89	0.7193	0.9248	0.5306	0.3003	0.9771
URL Embedding	URL-Embedded vs. De-Identified Resume (HRu1 vs. HRi1)	86	<i>0.0025</i>	0.0000	0.0000	<i>0.0002</i>	0.2214
URL Embedding	URL-Embedded Resume vs. LinkedIn (HRu1 vs. HL1)	83	0.7352	0.0000	0.3603	<i>0.0068</i>	0.7167

Table 12 continued

Facet	Input Versions	N	O	C	E	A	S
Source Context	De-Identified Resume vs. LinkedIn (HRi1 vs. HL1)	84	<i>0.0077</i>	0.3997	0.0000	0.1730	0.6718
Source Context	Original Resume vs. LinkedIn (HRo1 vs. HL1)	84	0.5300	<i>0.0003</i>	0.0001	0.0221	0.4553
Source Context	Original Resume vs. Twitter (HRo1 vs. HT1)	20	0.0121	0.0826	0.8983	<i>0.0020</i>	<i>0.0010</i>
Source Context	LinkedIn vs. Twitter (HL1 vs. HT1)	18	<i>0.0023</i>	<i>0.0047</i>	<i>0.0007</i>	<i>0.0047</i>	<i>0.0007</i>
Immediate Rep.	De-Identified Resume back-to-back (HRi2 vs. HRi3)	89	0.1797	0.3173	0.3173	0.6547	0.6547
Algorithm-Time	De-Identified Resume 31 days apart (HRi1 vs. HRi2)	89	<i>0.0071</i>	0.5314	0.2540	0.0516	0.2424
Participant-Time	LinkedIn 7–9 months apart (HL1 vs. HL2)	88	0.6487	0.0000	0.9615	<i>0.0072</i>	0.6011
Participant-Time	Twitter 7–9 months apart (HT1 vs. HT2)	21	N/A	N/A	N/A	N/A	N/A

Bold highlighting indicates value below Bonferroni-corrected threshold based on α of 0.05. Italic indicates p-value below Benjamini–Hochberg corrected threshold and above Bonferroni-corrected threshold. “N/A” values reflect experiments where there was zero change across the facet. Results are discussed in Sects. 5.3, 5.4, 5.5, 5.6, and 5.7

each measured on a scale from 0 to 10, so we divide the DiSC L1 distances by 40. Because *Crystal* constrains their DiSC scores to sum to 100, the maximum possible L1 change is 200, and we therefore use a normalization constant of 200. **Subgroup stability** We use demographic information provided in our survey to estimate rank-order stability, locational stability, and normalized L1 distance within subgroups defined by gender and primary language. With only 94 participants, we lacked the statistical power to perform statistical analysis on the smaller subgroups (e.g. birth country, race).

B Additional results

B.1 Inclusion of LinkedIn URL in resume

We discovered locational differences deemed significant by the Bonferroni threshold in *Dominance* (de-identified median 6.90, URL-embedded median 5.65; Wilcoxon $p < 10^{-6}$), *Big Five Conscientiousness* (de-identified median 5.60, URL-embedded median 6.17; Wilcoxon $p = 2.1 \times 10^{-5}$), and *Extraversion* (de-identified median 4.14, URL-embedded median 6.38; Wilcoxon $p < 10^{-6}$). Under the more liberal Benjamini–Hochberg standard, there were also significant locational differences in DiSC *Calculativeness* (de-identified median 7.50, URL-embedded median 8.00; Wilcoxon $p = 4.7 \times 10^{-3}$), *Openness* (de-identified median 6.14, URL-embedded

median 5.90; Wilcoxon $p = 2.5 \times 10^{-3}$), *Steadiness* (de-identified median 5.00, URL-embedded median 5.60; Wilcoxon $p = 4.8 \times 10^{-4}$), and *Agreeableness* (de-identified median 5.56, URL-embedded median 6.07; Wilcoxon $p = 1.6 \times 10^{-4}$).

Correlations between scores derived from LinkedIn profiles and from URL-embedded resumes ranged from 0.156 (*Dominance*) to 0.702 (*Emotional Stability*), and there was a significant difference in the medians of Big Five *Conscientiousness* (LinkedIn 5.72, resume 6.19; Wilcoxon $p = 4.3 \times 10^{-5}$), per the Bonferroni-adjusted threshold. Under Benjamini–Hochberg correction, the differences in *Dominance* (LinkedIn median 4.90, resume median 5.60; Wilcoxon $p = 6.6 \times 10^{-3}$) and *Agreeableness* (LinkedIn median 5.81, resume median 6.06; Wilcoxon $p = 6.8 \times 10^{-3}$) were significant as well. We predicted higher correlations under the embedding hypothesis, but a four month gap in algorithm-time as well as participant-time is likely to degrade the correlations significantly. Still, LinkedIn scores are more highly correlated with URL-embedded resumes than they are with de-identified resumes. Although instability due to algorithm-time is not guaranteed to increase monotonically with chronological time, this finding holds slightly more weight given that there were two more weeks of time between the LinkedIn and URL-embedding resume scoring. We also find that scores from URL-embedded resumes correlate slightly better with those from LinkedIn (generated four months earlier) than they do with those from de-identified resumes (generated just 2 weeks earlier).

B.2 Source context

Comparing de-identified resumes to LinkedIn profiles in *Humantic AI*, we found significant locational differences under Bonferroni in *Dominance* (LinkedIn median 4.85, resume median 6.85; Wilcoxon $p < 10^{-6}$) and *Extraversion* (LinkedIn median 6.44, resume median 4.06; Wilcoxon $p < 10^{-6}$), and under Benjamini–Hochberg in *Steadiness* (LinkedIn median 5.30, resume median 5.00; Wilcoxon $p = 1.3 \times 10^{-3}$) and *Openness* (LinkedIn median 6.01, resume median 6.14; Wilcoxon $p = 7.7 \times 10^{-3}$).

When original resumes were compared to LinkedIn profiles in *Humantic AI*, we observed significant locational differences under Bonferroni in *Dominance* (LinkedIn median 4.85, resume median 5.95; Wilcoxon $p = 7 \times 10^{-6}$) and *Extraversion* (LinkedIn median 6.44, resume median 5.75; Wilcoxon $p = 6.9 \times 10^{-5}$), and significant locational differences under Benjamini–Hochberg in *Influence* (LinkedIn median 4.60, resume median 4.85; Wilcoxon $p = 5.0 \times 10^{-3}$) and Big Five *Conscientiousness* (LinkedIn median 5.73, resume median 5.98; Wilcoxon $p = 2.8 \times 10^{-4}$). Although there was not any significant locational instability for *Agreeableness* overall, we found that for non-native English speakers, the median *Agreeableness* score on resumes (5.99) was significantly different under Benjamini–Hochberg ($p = 6.1 \times 10^{-3}$) from the median score on LinkedIn (5.63).

Comparing *Humantic AI* scores from Twitter to those from original resumes, we find correlations ranging from -0.521 (*Dominance*) to 0.232 (Big Five *Conscientiousness*). We easily avoid the issue of algorithm-time by using original resumes, which were run the same day as Twitter. None of the original resumes contain references to par-

ticipants' Twitter accounts, and furthermore we did not find evidence of linkage with Twitter profiles, so we need not worry about data leakage in this case. A major caveat to this result is the small sample size ($N = 20$). Although the locational differences were insignificant when compared to the Bonferroni-corrected threshold, the Benjamini–Hochberg correction found significant locational differences in *Agreeableness* (resume median 6.37, Twitter median 3.32; Wilcoxon $p = 2.0 \times 10^{-3}$) and *Emotional Stability* (resume median 5.42, Twitter median 7.97; Wilcoxon $p = 1.0 \times 10^{-3}$). Although there was not any significant locational instability for *Openness* overall, we found that for male participants, the median *Openness* score on resumes (5.71) was significantly different under Benjamini–Hochberg ($p = 6.1 \times 10^{-3}$) from the median score on Twitter (8.50).

Finally, we compare the *Humantic AI* scores from LinkedIn and Twitter. Again we have a small sample size ($N = 18$), however the results are striking. Only one of the correlations is positive (*Influence*, $r = 0.020$), and the others are as low as -0.433 (*DiSC Calculativeness*). Again there are no significant locational differences under Bonferroni, but using the Benjamini–Hochberg correction we find significant differences in *Openness* (LinkedIn median 5.82, Twitter median 8.16; Wilcoxon $p = 2.3 \times 10^{-3}$), *Big Five Conscientiousness* (LinkedIn median 5.77, Twitter median 7.16; Wilcoxon $p = 4.7 \times 10^{-3}$), *Extraversion* (LinkedIn median 6.80, Twitter median 4.72; Wilcoxon $p = 6.7 \times 10^{-4}$), *Agreeableness* (LinkedIn median 6.32, Twitter median 3.32; Wilcoxon $p = 4.7 \times 10^{-3}$), and *Emotional Stability* (LinkedIn median 4.86, Twitter median 7.97; Wilcoxon $p = 6.7 \times 10^{-4}$). Although there was not any significant locational instability for *Dominance* overall, we found that for male participants, the median *Dominance* score on LinkedIn (4.30) was significantly different under Benjamini–Hochberg ($p = 2.0 \times 10^{-3}$) from the median score on Twitter (6.90). Participant-time and algorithm-time are both guaranteed to be constant in this experiment, as profiles were generated on the same day.

Complete experimental results for *Humantic AI* are listed in Tables 8, 9, 11, and 12.

B.3 Algorithm time

Figure 8 shows that substandard sub-group correlations result from two participants whose resumes were scored very differently by *Humantic AI* a month apart; we also note that the lack of immediate reproducibility we observed in *Humantic AI* did not affect these two particular individuals. We did not find any significant locational differences across algorithm-time using the Bonferroni correction, but under Benjamini–Hochberg we found significant differences in *Openness*, where the median decreased from 6.15 to 6.13 over the course of a month (Wilcoxon $p = 7.1 \times 10^{-3}$).

B.4 Participant time

Built into the substandard correlations across participant-time in *Humantic AI* LinkedIn runs is the corrosive effect of 7–9 months of participant-time; this helps to explain, but does not justify, the unacceptably low test-retest reliability.

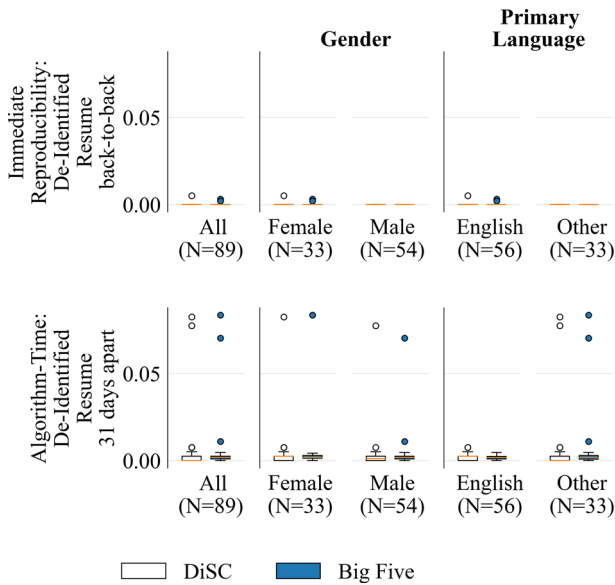


Fig. 8 Normalized L1 distances between *Humantic AI* DiSC and Big Five scores produced from identical resumes scored at different points in time

Under Bonferroni correction, we found the following significant difference in *Humantic AI* LinkedIn across 7–9 months of participant-time: Big Five *Conscientiousness* scores, with the median increasing from 5.72 to 6.17 (Wilcoxon $p = 4 \times 10^{-6}$). Under Benjamini–Hochberg we also found a significant difference in *Agreeableness*, where the median increased from 5.81 to 5.99 (Wilcoxon $p = 7.2 \times 10^{-3}$). Complete experimental results for *Humantic AI* are listed in Tables 8, 9, 11, and 12.

References

- Ajunwa I (2021) An Auditing Imperative for Automated Hiring Systems. *Harvard J Law & Tech* 34(2):80
- Araujo L, Saltelli A, Schnepf SV (2017) Do PISA data justify PISA-based education policy? *Inter J Comparative Educ Dev* 19(1):20–34
- Bandy J (2021) Problematic machine behavior: A systematic literature review of algorithm audits. *Proc. ACM Hum. Comput. Interact.* 5(CSCW1):1–34
- Barocas S, Selbst AD (2016) Big Data’s Disparate Impact. *Calif Law Rev* 671(104):671–732
- Bendick M (2007) Situation Testing for Employment Discrimination in the United States of America. *Horizons strategiques* 5(3):17–39
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Roy Stat Soc: Ser B (Methodol)* 57(1):289–300
- Blau FD, Brummund P, Liu AY-H (2013) Trends in Occupational Segregation by Gender 1970–2009: Adjusting for the Impact of Changes in the Occupational Coding System. *Demography* 50(2):471–494
- Bogen M, Rieke A (2018) Help Wanted: An Exploration of Hiring Algorithms. Equity and Bias, Technical report, Upturn
- Brown S, Davidovic J, Hasan A (2021) The algorithm audit: Scoring the algorithms that score us. *Big Data Soc* 8(1):1–8
- California Civil Code (2018). Title 1.81.5. California Consumer Privacy Act of 2018







- Cardinet J, Tourneur Y, Allal L (1976) The Symmetry of Generalizability Theory: Applications to Educational Measurement. *J Educ Meas* 13(2):119–135
- Carmines E, Zeller R (1979) Reliability and Validity Assessment. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States of America
- Chamorro-Premuzic T, Winsborough D, Sherman RA, Hogan R (2016) New Talent Signals: Shiny New Objects or a Brave New World? *Ind Organ Psychol* 9(3):621–640
- Chen L, Ma R, Hannák A, Wilson C (2018) Investigating the Impact of Gender on Rank in Resume Search Engines. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp 1–14. ACM
- Dastin J (2018) Amazon scraps secret AI recruiting tool that showed bias against women. Reuters
- Datta A, Tschantz MC, Datta A (2015) Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *Proc Privacy Enhancing Technol* 2015(1):92–112
- De-Arteaga M, Romanov A, Wallach HM, Chayes JT, Borgs C, Chouldechova A, Geyik SC, Kenthapadi K, Kalai AT (2019) Bias in bios: A case study of semantic representation bias in a high-stakes setting. In: danah boyd and Morgenstern, J. H., editors, Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, pp 120–128. ACM
- Emre M (2018) The Personality Brokers: The Strange History of Myers-Briggs and the Birth of Personality Testing, 1st edn. Doubleday, New York
- Equal Employment Opportunity Commission (EEOC), Civil Service Commission, Department of Labor, and Department of Justice (1978). Uniform guidelines on employee selection procedures. Federal Register
- European Commission (2021) Better regulation toolbox. https://ec.europa.eu/info/law/law-making-process/planning-and-proposing-law/better-regulation-why-and-how/better-regulation-guidelines-and-toolbox/better-regulation-toolbox_en; Accessed on 07/29/2022
- Fjeld J, Achten N, Hilligoss H, Nagy A, Sri Kumar M (2020) Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. SSRN Electronic Journal
- Galli A, Giampietro M, Goldfinger S, Lazarus E, Lin D, Saltelli A, Wackernagel M, Müller F (2016) Questioning the Ecological Footprint. *Ecol Ind* 69:224–232
- Goodstein LD, Lanyon RI (1999) Applications of Personality Assessment to the Workplace: A Review. *J Bus Psychol* 13(3):32
- Guion RM, Gottier RF (1965) Validity Of Personality Measures In Personnel Selection. *Personnel Psychology*, 18(2):135–164. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1744-6570.1965.tb00273.x>
- Guzzo RA, Fink AA, King E, Tonidandel S, Landis RS (2015) Big Data Recommendations for Industrial-Organizational Psychology. *Ind Organ Psychol* 8(4):491–508
- Hagendorff T (2020) The Ethics of AI Ethics: An Evaluation of Guidelines. *Mind Mach* 30(1):99–120
- Hannák A, Wagner C, Garcia D, Misllove A, Strohmaier M, Wilson C (2017) Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, pp 1914–1933, Portland Oregon USA. ACM
- Hegewisch A, Liepmann H, Hayes J, Hartmann H (2010) Separate and Not Equal? Gender Segregation in the Labor Market and the Gender Wage Gap. Technical report, Institute for Women’s Policy Research
- Hough LM, Eaton NK, Dunnette MD, Kamp JD, McCloy RA (1990) Criterion-related validities of personality constructs and the effect of response distortion on those validities. *J Appl Psychol* 75(5):581–595
- Hurtz GM, Donovan JJ (2000) Personality and job performance: The Big Five revisited. *J Appl Psychol* 85(6):869–879
- Kelly-Lyth A (2020) Challenging Biased Hiring Algorithms. SSRN Scholarly Paper ID 3744248. Social Science Research Network, Rochester, NY
- Kim PT (2017) Data-Driven Discrimination at Work. *William & Mary Law* 58:81
- Köchling A, Wehner MC (2020) Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of hr recruitment and hr development. *Bus Res* 13(3):795–848
- Koshiyama A, Kazim E, Treleaven P, Rai P, Szpruch L, Pavey G, Ahamat G, Leutner F, Goebel R, Knight A, Adams J, Hitrova C, Barnett J, Nachev P, Barber D, Chamorro-Premuzic T, Klemmer K, Gregorovic M, Khan S, Lomas E (2021) Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI. ML and Associated Algorithms, Technical report

- Langenkamp M, Costa A, Cheung C (2020) Hiring Fairly in the Age of Algorithms. [arXiv:2004.07132](https://arxiv.org/abs/2004.07132) [cs]. [arXiv: 2004.07132](https://arxiv.org/abs/2004.07132)
- Lo Piano S, Robinson M (2019) Nutrition and public health economic evaluations under the lenses of post normal science | Elsevier Enhanced Reader. *Futures*, 112
- Lussier K (2018) Temperamental workers: Psychology, business, and the Humm-Wadsworth Temperament Scale in interwar America. *Hist Psychol* 21(2):79
- Meinert D (2015) What Do Personality Tests Really Reveal? *HR Magazine*, SHRM. <https://www.shrm.org/hr-today/news/hr-magazine/pages/0615-personality-tests.aspx>; accessed on 07/29/2022
- Metcalf J, Moss E, Watkins EA, Singh R, Elish MC (2021) Algorithmic impact assessments and accountability: The co-construction of impacts. In: Elish MC, Isaac W, Zemel RS (eds) *FACCT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, pp 735–746
- Morgeson FP, Campion MA, Dipboye RL, Hollenbeck JR, Murphy K, Schmitt N (2007) Reconsidering the use of personality tests in personnel selection contexts. *Pers Psychol* 60(3):683–729
- Morrow JR, Jackson Aw (1993) How Significant is Your Reliability? *Res Q Exerc Sport* 64(3):352–355
- Mueller RO, Knapp TR (2018) Reliability and Validity. In: *The Reviewer's Guide to Quantitative Methods in the Social Sciences*. Routledge, 2 edition
- Mökander J, Morley J, Taddeo M, Floridi L (2021) Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. *Sci Eng Ethics* 27(4):44
- New York City Council (2021). Local Law 144 to amend the administrative code of the city of New York, in relation to automated employment decision tools
- Nunnally JC, Bernstein IH (1994) *Psychometric Theory*, 3rd edn. McGraw Hill, New York, NY
- Oala L, Fehr J, Gilli L, Balachandran P, Leite AW, Ramirez SC, Li DX, Nobis G, Alvarado EAM, Jaramillo-Gutierrez G, Matek C, Shroff A, Kherif F, Sanguinetti B, Wiegand T (2020) ML4H auditing: From paper to practice. In: Alsentzer E, McDermott MBA, Falck F, Sarkar SK, Roy S, Hyland SL, editors, *Machine Learning for Health Workshop, ML4H@NeurIPS*, volume 136 of *Proceedings of Machine Learning Research*, pages 280–317. PMLR
- ORCAA (2020) Description of Algorithmic Audit: Pre-built Assessments. Technical report
- Raghavan M, Barocas S, Kleinberg J, Levy K (2020) Mitigating bias in algorithmic hiring: evaluating claims and practices. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp 469–481. ACM
- Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Hildebrandt M, Castillo C, Celis LE, Ruggieri S, Taylor L, Zanfir-Fortuna G (eds) *FAT* '20: Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, pp 33–44
- Razavi S, Jakeman A, Saltelli A, Prieur C, Iooss B, Borgonovo E, Plischke E, Lo Piano S, Iwanaga T, Becker W, Tarantola S, Guillaume JHA, Jakeman J, Gupta H, Melillo N, Rabitti G, Chabridon V, Duan Q, Sun X, Smith S, Sheikholeslami R, Hosseini N, Asadzadeh M, Puy A, Kucherenko S, Maier HR (2021) The Future of Sensitivity Analysis: An essential discipline for systems modeling and policy support. *Environ Model Software* 137:104954
- Riksrevisjonen (2020) Auditing machine learning algorithms: Report by the Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK. <https://www.auditingalgorithms.net/>; Accessed on 07/29/2022
- Robertson RE, Jiang S, Joseph K, Friedland L, Lazer D, Wilson C (2018) Auditing Partisan Audience Bias within Google Search. *Proc ACM Human-Comput Interaction* 2(CSCW):1–22
- Saltelli A, Bamber G, Bruno I, Charters E, Di Fiore M, Didier E, Nelson Espeland W, Kay J, Lo Piano S, Mayo D, Pielke R Jr, Portaluri T, Porter TM, Puy A, Rafols I, Ravetz JR, Reinert E, Sarewitz D, Stark PB, Stirling A, van der Sluijs J, Vineis P (2020) Five ways to ensure that models serve society: a manifesto. *Nature* 582(7813):482–484
- Saltelli A, Lo Piano S (2017) Problematic Quantifications: a Critical Appraisal of Scenario Making for a Global Sustainable Food Production. *Food Ethics* 1(2):173–179
- Sandvig C, Hamilton K, Karahalios K, Langbort C (2014) Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. In: *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*, p 23, Seattle, WA, USA
- Schellmann H, Strong J, Siegel I (2021a) Hired by an algorithm. In: *Machines We Trust* podcast series, MIT Technology Review. issued: 2021-06-23
- Schellmann H, Strong J, Siegel I (2021b) Want a job? The AI will see you now. In: *Machines We Trust* podcast series, MIT Technology Review. issued: 2021-07-07

- Schmidt FL, Le H, Ilies R (2003) Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychol Methods* 8(2):206
- Schmitt N, Gooding RZ, Noe RA, Kirsch M (1984) Metaanalyses of Validity Studies Published Between 1964 and 1982 and the Investigation of Study Characteristics. *Personnel Psychology*, 37(3):407–422. [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1744-6570.1984.tb00519.x](https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1744-6570.1984.tb00519.x)
- Science Advice for Policy by European Academies (SAPEA) (2019) Making sense of science for policy under conditions of complexity and uncertainty. Science Advice for Policy by European Academies, DE
- Scroggins WA, Thomas SL, Morris JA (2008) Psychological Testing in Personnel Selection, Part I: A Century of Psychological Testing. *Public Personnel Manag* 37(1):99–109
- Sharma S, Henderson J, Ghosh J (2020) CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp 166–172. ACM
- Shneiderman B (2020) Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Trans Interactive Intell Syst* 10(4):1–31
- Sloane M (2021) The Algorithmic Auditing Trap. <https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d>; Accessed 07/29/2022
- Sloane M, Moss E, Chowdhury R (2022) A silicon valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability. *Patterns* 3(2):100425
- Stark L, Hutson J (2021) Physiognomic Artificial Intelligence. SSRN Electronic Journal
- Stoyanovich J (2021) Hiring and AI: Let Job Candidates Know Why They Were Rejected. *The Wall Street Journal*
- Sühr T, Hilgard S, Lakkaraju H (2021) Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp 989–999. ACM
- The European Commission (2021) Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts
- Turney P (1995) Technical note: Bias and the quantification of stability. *Mach Learn* 20(1–2):23–33
- VanderWeele TJ, Mathur MB (2019) Some desirable properties of the Bonferroni correction: Is the Bonferroni correction really so bad? *Am J Epidemiol* 188(3):617–618
- Vecchione B, Levy K, Barocas S (2021) Algorithmic auditing and social justice: Lessons from the history of audit studies. In: EAAMO 2021: ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, Virtual Event, USA, October 5 - 9, 2021, pp 19:1–19:9. ACM
- Venkataadri G, Andreou A, Liu Y, Mislove A, Gummadi KP, Loiseau P, Goga O (2018) Privacy Risks with Facebook’s PII-Based Targeting: Auditing a Data Broker’s Advertising Interface. In: 2018 IEEE Symposium on Security and Privacy (SP), pp 89–107. ISSN: 2375-1207
- Weber L, Dvoskin E (2014) Are Workplace Personality Tests Fair? *Wall Street Journal*
- Wilson C, Ghosh A, Jiang S, Mislove A, Baker L, Szary J, Trindel K, Polli F (2021) Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp 666–677. ACM
- Xue S, Yurochkin M, Sun Y (2020) Auditing ML Models for Individual Bias and Unfairness. In: International Conference on Artificial Intelligence and Statistics, pp 4552–4562. PMLR. ISSN: 2640-3498

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Alene K. Rhea^{1,2}  · Kelsey Markey^{1,2}  · Lauren D’Arinzo^{1,2,3}  ·
 Hilke Schellmann⁴ · Mona Sloane²  · Paul Squires⁵ · Falaah Arif Khan^{1,2}  ·
 Julia Stoyanovich^{1,2,6} 

Alene K. Rhea
alene@nyu.edu

Kelsey Markey
kelseymarkey@nyu.edu

Lauren D' Arinzo
lauren.darinzo@nyu.edu

Hilke Schellmann
hilke.schellmann@nyu.edu

Mona Sloane
mona.sloane@nyu.edu

Paul Squires
ps2937@nyu.edu

Falaah Arif Khan
fa2161@nyu.edu

- 1 Center for Data Science, New York University, New York, USA
- 2 Center for Responsible AI, Tandon School of Engineering, New York University, Brooklyn, USA
- 3 The MITRE Corporation, Bedford, MA, USA
- 4 Arthur L. Carter Journalism Institute, New York University, New York, USA
- 5 Department of Psychology, Arts & Science, New York University, New York, USA
- 6 Computer Science & Engineering, Tandon School of Engineering, Brooklyn, USA