Resume Format, LinkedIn URLs and Other Unexpected Influences on AI Personality Prediction in Hiring: Results of an Audit

Alene K. Rhea
New York University
New York, USA
alene@nyu.edu

Kelsey Markey
New York University
New York, USA
kelseymarkey@nyu.edu

Lauren D'Arinzo*
New York University
New York, USA
lauren.darinzo@nyu.edu

Hilke Schellmann
New York University
New York, USA
hilke.schellmann@nyu.edu

Mona Sloane
New York University
New York, USA
mona.sloane@nyu.edu

Paul Squires
New York University
New York, USA
ps2937@nyu.edu

Julia Stoyanovich†
New York University
New York, USA
stoyanovich@nyu.edu

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 reliability of such systems using stability of the outputs they produce, noting that reliability is a necessary, but not a sufficient, condition for validity. We develop a methodology for an external audit of stability of algorithmic personality tests, and instantiate this methodology in an audit of two systems, Humantic AI and Crystal. Rather than challenging or affirming the assumptions made in psychometric testing — that personality traits are meaningful and measurable constructs, and that they are indicative of future success on the job — we frame our methodology around testing the underlying assumptions made by the vendors of the algorithmic personality tests themselves.

In our audit of Humantic AI and Crystal, we find that both systems show substantial instability on key facets of measurement, and so cannot be considered valid testing instruments. For example, Crystal frequently computes different personality scores if the same resume is given in PDF vs. in raw text, violating the assumption that the output of an algorithmic personality test is stable across job-irrelevant input variations. Among other notable findings is evidence of persistent — and often incorrect — data linkage by Humantic AI.

An open-source implementation of our auditing methodology, and of the audits of Humantic AI and Crystal, is available at https://github.com/DataResponsibly/hiring-stability-audit.

CCS CONCEPTS

• **Social and professional topics** → **Socio-technical systems**; **Employment issues**; *Testing, certification and licensing*; **Automation**.

KEYWORDS

algorithm audit, validity, stability, reliability, hiring, personality

ACM Reference Format:
Alene K. Rhea, Kelsey Markey, Lauren D'Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, and Julia Stoyanovich. 2022. Resume Format, LinkedIn URLs and Other Unexpected Influences on AI Personality Prediction in Hiring: Results of an Audit. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3514094.3534189

1 INTRODUCTION

AI-based automated hiring systems are seeing ever broader use. 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. In this paper, we focus on automated pre-hire assessment, as some of the fastest-developing of all high-stakes uses of AI [27]. Reports of algorithmic hiring systems acting in ways that are discriminatory or unreliable abound [3, 6, 12, 13, 18, 60]. For example, an automated phone interview tool was found to produce high "English competency" scores even when the candidate spoke in German or Chinese [51], undermining the tool's *validity*.

*Also with The MITRE Corporation. Affiliation with 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 22-0647.©2022 The MITRE Corporation. All rights reserved.
†Contact author.

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08…$15.00
https://doi.org/10.1145/3514094.3534189

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 [15, 32, 58], and because algorithmic personality tests are rarely validated by third-parties [50]. Our approach is to develop a methodology for an *external audit of stability* of predictions made by algorithmic personality tests. Stability is closely related to the psychometric concept of reliability, which is a prerequisite of validity (see Section 2.1 for details). Crucially, we frame our methodology around *testing the underlying assumptions made by the vendors of the algorithmic personality tests themselves* [59].

In this paper, we make the following contributions: We start with an overview of the key literature on psychometric testing applied to hiring and on algorithm auditing with a focus on hiring (Section 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. We then develop a quantitative methodology, informed by psychometric theory and sociology, to audit the stability of algorithms that predict personality for use in hiring (Section 3). Figure 1 gives an overview of our proposed methodology.

We instantiate this methodology in an audit of two systems, `Humantic AI` and `Crystal`, over a dataset of job applicant profiles collected through an IRB-approved study (Section 4). We selected these systems because they accept easily-manipulated textual features as input, allow multiple input types, and produce quantitative personality traits as output. Additionally, these systems have substantial presence in the algorithmic hiring market: On their website, `Humantic AI` reports that it is used by Apple, PayPal and McKinsey, while `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.

The results of our audit are summarized in Table 1. We find that both `Humantic AI` and `Crystal` show substantial instability on important facets of measurement, and so cannot be considered valid testing instruments. For example, 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, violating the assumption that the output is stable across job-irrelevant variations in the input. We also found that `Humantic AI` creates a persistent (and sometimes incorrect!) linkage between an email address and a LinkedIn URL that appear in a resume, and then silently disregards resume information when computing the personality score.

We discuss the results and limitation of our work in Section 5, and conclude in Section 6.

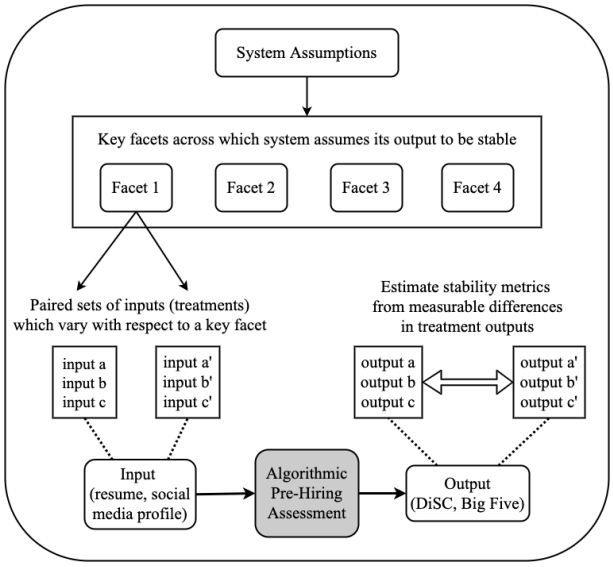

Figure 1: Framework for auditing methodology (Sec. 3)

Table 1: Summary of stability results for `Crystal` and `Humantic AI`, with respect to facets of measurement from Section 4.2. "✓" indicates sufficient rank-order stability ($r \geq 0.90$) and sufficient locational stability ($p \geq \alpha_{\text{Benjamini-Hochberg}}$) in all traits, "✗" indicates 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.

Facet	`Crystal`	`Humantic`	Details
Resume file format	✗	✓	Sec. 4.5.4
LinkedIn URL in resume	?	✗	Sec. 4.5.5
Source context	✗	✗	Sec. 4.5.6
Algorithm-time / immediate	✓	✓	Sec. 4.5.7
Algorithm-time / 31 days	✓	✗	Sec. 4.5.7
Participant-time / LinkedIn	✗	✗	Sec. 4.5.8
Participant-time / Twitter	N/A	✓	Sec. 4.5.8

2 BACKGROUND AND RELATED WORK

2.1 Personality Testing for Hiring

Since the early 1900s, personnel selection practices have relied on the use of psychometric instruments such as personality tests to identify promising candidates [55]. And although this practice is both longstanding and wide-spread [33], 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 [55]. 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" [21]. Several other surveys would come to the same conclusion in the following decades [25, 53], yet, Human Resources (HR) professionals continued to use personality testing for hiring [55]. 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. The use of a traditional personality test in personnel selection relies on the following assumptions:

- personality traits being measured are meaningful constructs;
- the test is a valid measurement instrument: it measures the traits it purports to measure; and
- 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 [8, 9]. *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 [37]. Reliability is a necessary (although not a sufficient) condition for validity [39]. 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. [37]. Each dimension across which measurements are compared is referred to as a "facet," such that we consider reliability with respect to some facet (e.g., time) that varies between measurements, while other facets (e.g., test location) are held constant [8]. Under Classical Test Theory (CTT), measurements can be decomposed into a true score and a measurement error [52]. The true score is the value of the underlying construct of interest, while measurement error can be further broken down across various experiment facets [52].

Reliability is usually measured and evaluated with correlations. Although a correlation coefficient of 0.80 is often cited as an acceptable threshold of reliability, Nunnally and Bernstein [39] 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, they advise that 0.90 should be the "bare minimum," and that 0.95 should be the "desirable standard."

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.* [22] caution that reliability and validity are "often overlooked yet critically important" in big-data applications of I-O psychology. 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

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.* [29]. Scholarly work on algorithm auditing acknowledges that auditing frameworks are inconsistent in terms of scope, methodology, and evaluation metrics [3, 7, 29, 43]. Much of the audit literature surrounding predictive hiring technology is primarily concerned with legal liability as laid out in the Uniform Guidelines on Employee Selection Procedures [28, 42, 66]. These guidelines, adopted by the US Equal Employment Opportunity Commission in 1978 [16], 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. As a result, audits of AI hiring systems are often specifically concerned with adverse impact [10, 41, 66]. It is often noted that avoiding liability is not actually sufficient to ensure an ethical system: a lack of adverse impact should be a baseline rather than the end goal [4, 41, 42, 66].

One of the contributions of our work is an audit framework specific to personality prediction systems used in the hiring domain, and a technical instantiation of this framework for two candidate screening systems. As we will discuss in Section 3, our methodology is specific to the domain and to the tool under study [59].

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 [23, 30, 34, 49, 62, 65, 66]. Of the works that do, some refer to this concept as "stability" [7, 29, 46, 59, 61], some refer to it as "reliability" [17, 38, 43, 57, 61], and some refer to it as "robustness" [10, 17, 38, 40, 41]. Bandy [3] forgoes specific terminology and simply refers to changes to input and output. This difference is more than terminological: stability relates to local numerical analyses, robustness refers 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.* [44] observe that sensitivity analysis is not a unified discipline, but is instead spread across many fields, and notes that lack of common terminology remains a barrier to unification. 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. (In contrast, if small changes in the input lead to large changes in the output, then the algorithm is considered unstable.) We adopt a psychometric definition of *reliability*, using it to guide how we measure stability. By considering algorithms within their sociotechnical 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 [63]. The 2020 manifesto on responsible modeling [47] underscores the importance of sensitivity analysis, and both the European Commission [11] and the European Science Academies [54] have called for sensitivity auditing in the policy domain. Sensitivity audits have also been applied in the domains of education [2], food security [48], public health [31], and sustainability [19]. We argue that algorithm auditors should consider stability among the critical metrics.

Our work is synergistic with two recent lines of work that contribute substantive quantitative methodologies for auditing algorithm stability. Xue *et al.* [67] introduce a suite of tools to study individual fairness in black-box models. Sharma *et al.* [56] offer 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 sociotechnical lens, borrowing metrics that are familiar to I-O psychologists.

Audit scope. Several recent algorithm audits focus on tools used in hiring. Wilson *et al.* [66] and O'Neil Risk Consulting and Algorithmic Auditing [41] each focus on tools for pre-employment assessment (i.e., candidate screening). Raghavan *et al.* [42] evaluate the public claims about bias made by the vendors of 18 such tools. Chen *et al.* [10] audit three resume search engines, Hannák *et al.* [24] audit two online freelance marketplaces, and De-Arteaga *et al.* [14] build and evaluate several classifiers that predict occupation from online bios. All of these studies focus primarily on bias and discrimination. 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 (executed through a collaboration between internal and external stakeholders), or external (where auditors are fully independent and do not work directly with vendors). Sloane *et al.* [59] 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 [1] argues for both internal and external auditing imperatives, with the latter ideally performed by a new certifying authority. Brown *et al.* [7] offer a flexible framework for external audits that centers on stakeholder interests. Bogen and Rieke [6] stress the importance of independent algorithm evaluations and place the burden on vendors and employers to be "dramatically" more transparent. Absent that transparency, external audits must be designed around what information is publicly available. We develop an external auditing methodology in this work.

3 AUDITING METHODOLOGY

In accordance with Sloane *et al.* [59], we frame our methodology around testing the underlying assumptions made by algorithmic personality tests within the hiring domain. Because algorithmic personality tests constitute a category of psychometric instrument, they are subject to assumptions made by the traditional instruments, laid out in Section 2.1. Validity of these tests is subject to the following non-exhaustive list of additional assumptions:

- **A1:** The output of an algorithmic personality test is stable across supported input types (e.g., PDF or text) and other job-irrelevant variations in the input, based on *parallel forms reliability* in psychometric testing (see Section 2.1).
- **A2:** The output of an algorithmic personality test is stable across input sources (e.g., resume or LinkedIn) that are treated as interchangeable by the vendor, based on *cross-situational consistency*.
- **A3:** The output of an algorithmic personality test on the same input is stable over time, based on *test-retest reliability*.

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.

Audit procedure. We now present a procedure, shown in Figure 1, to assess the stability of algorithmic personality tests in hiring, inspired by the auditing framework of Brown *et al.* [7]. Our method requires numeric output: a single personality measure or a vector.

1. **Collect preliminary information** to describe the sociotechnical context in which the system operates, and detail the system's inputs and outputs.
2. **Identify key facets of measurement** across which the system assumes its outputs to be stable, based on validity assumptions.
3. **Collect or create an input corpus** that is representative of the tool's intended context of use. Perturb the input across the features that correspond to each facet of measurement, while keeping other features fixed, generating two *treatments* for assessing stability of each facet.
4. **Estimate stability across each key facet** by querying the system with the treatments for that facet. Record system outputs and compare them to assess stability.

Stability metrics. The following metrics can be used, but other metrics may also be applicable:

- **Rank-order stability.** Reliability of psychometric instruments is measured with correlations (see Section 2.1). Morrow and Jackson [36] make a convincing argument against providing significance levels for reliability correlations. Instead, we compare estimated correlations to the "bare minimum" of 0.90 and the "desirable standard" of 0.95, as proposed by Nunnally and Bernstein [39].
- **Locational stability.** Locational stability is distinct from rank-order stability; neither one implies the other. If a system allows users to compare output across a facet, then we should also assess locational stability across that facet, to determine whether one treatment generally yields higher overall scores. Choose an appropriate hypothesis test (e.g., paired t-test to compare treatment means, or the Wilcoxon signed-rank test to check whether the median of the paired differences is significantly different than 0). Account for multiple hypothesis testing by adjusting the significance threshold using Bonferroni or Benjamini-Hochberg correction. 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 [64].

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

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} [5].

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

- **Total change.** To quantify the total change across a facet, select a distance measure and a normalization procedure that are appropriate for the type of system output.
- **Subgroup stability.** If subject-level demographic data is available, compute metrics for rank-order stability, locational stability, and total change within each demographic group.

Additional details about our methodology are presented in Rhea *et al.* [45], where we discuss our extensible stability audit framework. Next, we will use this methodology to audit two algorithmic personality tests, `Humantic AI` and `Crystal`. An open-source implementation of our auditing framework, including an implementation of the audits of `Humantic AI` and `Crystal`, is available at https://github.com/DataResponsibly/hiring-stability-audit.

4 AUDITS OF HUMANTIC AI AND CRYSTAL

4.1 Preliminary Information

Systems of interest. We assess stability of two automated text-based employee screening systems provided by vendors `Humantic AI` and `Crystal`. 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).[1] Although official DiSC documentation states that C represents "Conscientiousness," `Humantic AI` states that C in DiSC stands for "Calculativeness."[2] Notably, although both `Humantic AI` and `Crystal` market DiSC as a rigorous psychology-based analysis methodology, scholarly work on DiSC in I-O psychology 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."[3]

The Big Five model is far better studied than DiSC, and its use in personnel selection is considered acceptable by some I-O psychologists [20, 26]). Still, the use of the Big Five in hiring is not without criticism. For example, Morgeson *et al.* [35] argue that "the validity of personality measures as predictors of job performance is often disappointingly low." 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."[4]

`Humantic AI` also outputs seven traits, which they call the "Behavioral Work Factors." We do not include these traits in our analysis. Finally, `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 feasible, we leave a treatment of stability in these textual profiles to future work.

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."[5] `Crystal` explains that their personality profiles are "predicted through machine learning and use text sample analysis and attribute analysis."[6] Neither company makes its training data publicly available, or discusses their data collection and selection methodology. Thus, 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%"[7] `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."[8] 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.

Sociotechnical context of use. 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 scoring and ranking candidates based on their personality profiles. `Crystal` assigns a "job fit" score to candidates, which is measured based on a comparison either to 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 these systems support are not fully automated. Human decision-makers — HR professionals — 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 highest-scoring 25 candidates. As these examples illustrate, the context of use of employee screening systems like `Humantic AI` and `Crystal` is crucial to actual outcomes.

[1] https://www.discprofile.com/what-is-disc/how-disc-works
[2] `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.
[3] https://www.discprofile.com/everything-disc/hiring
[4] https://app.humantic.ai/#/candidates
[5] https://api.humantic.ai/
[6] https://www.crystalknows.com/blog/crystal-accuracy
[7] https://api.humantic.ai/
[8] https://www.crystalknows.com/blog/crystal-accuracy

4.2 Key Facets of Measurement

We identify the following key facets across which Humantic AI and Crystal operationalize reliability, as discussed in Section 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. 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.

4.3 Creation of the Input Corpus

Primary data collection. We conducted an IRB-approved human subjects research study to seed the input corpus for the audit. 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. 60% of participants identified as male, 38% as female, and 2 (2%) as non-binary. Their ages ranged from 21-40 with a mean of 26. 60% of our sample identified as Asian, 25% as White, 5% as Hispanic or Latino, 3% as Black or African American, and 4% identified as two or more races. 1% declined to identify their race. 37% of participants were born in India, 30% were born in the US, 13% were born in China, and 20% were born elsewhere. 64% reported that English was their primary language.

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 fake 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 substantiate that Humantic AI operationalizes assumption **A2** of cross-situational consistency (Section 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 2 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 Section 4.5.2.

Generating treatments for each facet. 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 2. Details of each set of score-generating model calls that use these resume versions, or social media links, are presented in Appendix A.1. We will explain how these versions are used as treatments in the stability experiments in Section 4.4.

Table 2: 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.

4.4 Estimating Stability across Key Facets

To measure stability, we conduct a series of local sensitivity analyses [44] to probe the sensitivity of predicted personality traits to facets of interest. To conduct this analysis, 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 $754. 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 [29, 59]. 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.

We performed the following experiments to test stability with respect to the key facets of measurement, described in Section 4.2:

- **Resume file format.** We tested sensitivity to file format by generating identical resumes in different formats. `Humantic AI` accepts PDF and Microsoft Word DOCX documents, so we compared the output from de-identified resumes in PDF format (Table 3 run ID HRi1) to those same resumes in DOCX format (Table 3 run ID HRd1). `Crystal` accepts PDF or raw text documents, so we compared the output from raw text resumes (Table 4 run ID CRr1) to those same resumes in PDF format (Table 4 run ID CRp1).
- **Inclusion of LinkedIn URL in resume.** For each participant who submitted both a resume and a LinkedIn profile, we compared their `Humantic AI` personality profile results from de-identified resumes (Table 3 run ID HRi1) to the same resumes with the hyperlinked LinkedIn URL added before the first character of the resume, in the form https://www.linkedin.com/in/ParticipantUsername (i.e., URL-embedded resumes, Table 3 run ID HRu1).
- **Source context.** We tested `Humantic AI`'s sensitivity to input source context by comparing output from participants' LinkedIn profiles (Table 3 run ID HL1), Twitter accounts (Table 3 run ID HT1), and resumes. Comparisons including resumes were repeated with original (Table 3 run ID HRo1), URL-embedded (Table 3 run ID HRou1), and de-identified resumes (Table 3 run ID HRi1). For `Crystal`, we compared output from PDF resumes (Table 4 run ID CRp1) to output from LinkedIn (Table 4 run ID CL1).
- **Algorithm-time / immediate.** We assessed the extent to which results from each system were immediately reproducible by inputting the same resume twice, consecutively. We compared de-identified resumes in `Humantic AI` (Table 3 run IDs HRi2 and HRi3), and raw text resumes in `Crystal` (Table 4 run IDs CRr2 and CRr3).
- **Algorithm-time / 31 days.** We also tested the sensitivity of scores to longer differences in algorithm-time, by comparing the output of identical resumes scored 31 days apart from one another. The same resume versions were used in this comparison as in the algorithm-time / immediate experiment: we used de-identified resumes in `Humantic AI` (Table 3 run IDs HRi1 and HRi2) and raw text resumes in `Crystal` (Table 4 run IDs CRr1 and CRr2).
- **Participant-time.** To test the effect of participant-time differences on outcomes, we generated two time-separated scores from participants' LinkedIn profiles (Table 3 run IDs HL1 and HL2; Table 4 run IDs CL1 and CL2). In `Humantic AI` we also generated two time-separated scores from participants' Twitter accounts (Table 3 run IDs HT1 and HT2). The time elapsed between the sets of scores ranged from seven to nine months in `Humantic AI`, and eight to ten months in `Crystal`. This test was performed on social media profiles rather than on resumes because participants naturally update their social media profiles, whereas accessing updated resumes would require a second round of primary data collection from study participants.

We attempted to isolate the key facet of interest in each experiment by keeping all other measurement facets constant across the pair of treatments. In some cases, this was not possible (e.g., measuring across participant-time on social media necessitates also measuring across algorithm-time; see Section 4.5.8). Additionally, we discovered problematic mechanisms in `Humantic AI` (i.e., imperfect immediate reproducibility and linkage between email addresses and LinkedIn accounts) only after performing initial experiments, at which time it was no longer feasible to re-run all experiments. We chose to prioritize the use of de-identified resumes at the expense of allowing variations in algorithm-time. The implications of this choice are discussed in Sections 4.5.5 and 4.5.6.

In what follows, we present audit results. See Appendix A.2 for details on specific stability metrics.

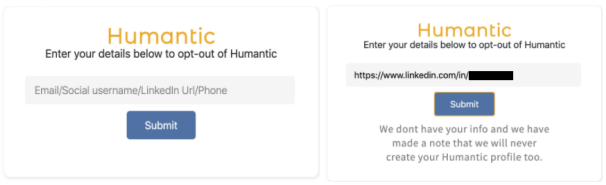

Figure 2: Screen shots of the `Humantic AI` "opt out" feature.

4.5 Results

4.5.1 Summary of experimental results. Table 1 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. We discuss experimental results in the remainder of this section, and present additional details in Appendix A.3.

4.5.2 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. Notably, the email address in question need not be associated with the LinkedIn profile, or even with the candidate. In one case, a participant listed contact information for references, and `Humantic AI` created a link between a reference's email address 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 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 Figure 2), this feature seems to be inoperable. Various forms of participant

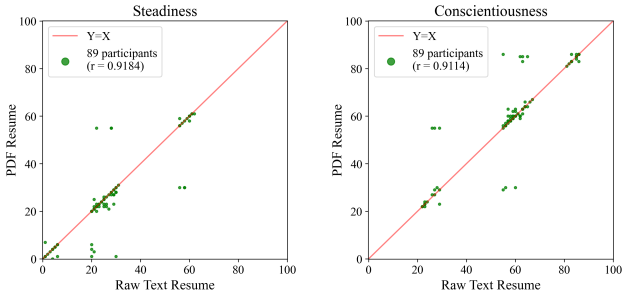

Figure 3: Comparison of `Crystal` output across the resume file format facet. Note evidence of discontinuous measurement in DiSC Steadiness and Conscientiousness, with some participants' scores moving between clusters with different file formats.

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 LinkedIns, even on different `Humantic AI` accounts.

4.5.3 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. This can lead to increased instability when a small change in input leads to a large change in output across the point of discontinuity. We observed evidence of this phenomenon in both Steadiness and Conscientiousness across all facets in `Crystal`, see Figure 3 for an example.

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 Dominance score was always 5, the median Influence score was always 10, the median Steadiness score was always 22 or 23, and the median Conscientousness score ranged from 59 to 62. This result is especially notable, considering that we observed rank-order instability in `Crystal`. (See Sections 4.5.4, 4.5.5, 4.5.6, 4.5.7, and 4.5.8.)

4.5.4 File format. We determined 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

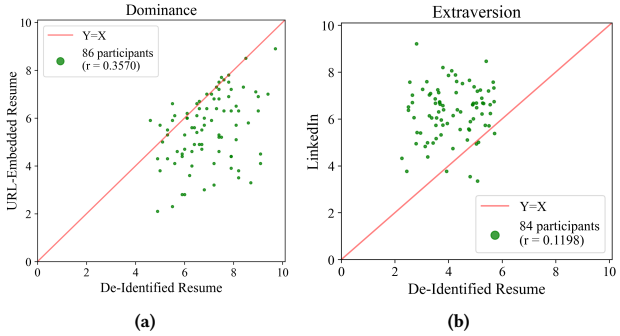

Figure 4: (a) `Humantic AI` Dominance scores from de-identified and URL-embedded resumes. (b) `Humantic AI` Extraversion scores produced by de-identified resumes and LinkedIn profiles.

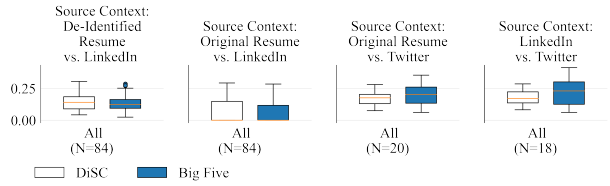

Figure 5: Normalized L1 distances between `Humantic AI` DiSC and Big Five scores produced from pairs of treatments that vary with respect to their input source.

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`.

4.5.5 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 (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 4(a) gives a representative example; see Appendix A.3.1 for complete results.

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.

4.5.6 Source context. `Humantic AI` and `Crystal` both displayed low stability across input sources. See Figure 5 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 Section 4.5.7, `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 A.3.2 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 original 84 resumes contained a 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 A.3.2 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 4(b) highlights some of these results. Appendix A.3.2 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.

4.5.7 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. One participant had changes in their Dominance and Influence scores (DiSC total normalized

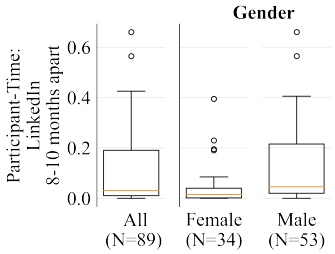

Figure 6: Normalized L1 distances between Crystal DiSC scores produced from LinkedIn profiles scored 8-10 months apart.

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 A.3.3 for additional details about this experiment.

4.5.8 Participant-time. Humantic AI scores on Twitter accounts showed no change over 7-9 months. However, 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 6 for cross-gender comparison of L1 distances between participant-time treatments. There was no significant locational instability across participant-time in Crystal. See Appendix A.3.4 for additional details about this experiment.

5 STUDY LIMITATIONS

In our audit we do not conduct stakeholder evaluations. Several audits and framework documents emphasize the importance of algorithmic impact assessment and stakeholder evaluations [7, 17, 41, 43, 44, 59]. Metcalf *et al.* [34] explain that an external audit must not stand in as an impact assessment.

Humantic AI often fails to produce profiles from inputs (see the discrepancies between number of inputs submitted and number of profiles produced in Table 3). This is especially common in Twitter profiles. By simply disregarding the failed inputs, we may be introducing some sampling bias into our results. Furthermore, such non-results themselves may exhibit problematic biases [41].

Although our audit considers various dimensions of reliability and stability, the analysis is not exhaustive. We analyze DiSC and Big Five scores, which claim to offer a quantitative measure of "personality." However, much of the advertising of both tools focuses on the profiles holistically, not just on the scores. Additionally, we evaluate the intermediate personality profile results and do 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, leaving critical questions about discrimination out of scope for this study.

6 CONCLUSIONS

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 of 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. We then developed a quantitative methodology, informed by psychometric theory and sociology, to audit the stability of black-box algorithms that predict personality for use in hiring, and instantiated it in an external audit of two systems, Humantic AI and Crystal. We found that both systems lack reliability across key facets of measurement, and concluded that they cannot be considered valid personality assessment instruments.

Our methodology can be used by employers to make informed purchasing and usage decisions, by legislators to guide regulation, and by job seekers to make informed decisions about disclosing their information to potential employers.

We demonstrated that stability, though often overlooked, 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 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.

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. 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.

7 ACKNOWLEDGEMENTS

This research was supported in part by NSF Awards No. 1934464, 1916505, and 1922658. We thank Dhara Mungra for her work on data collection and preliminary analysis, Daphna Harel and Joshua R. Loftus for their advice on statistical methods, and Falaah Arif Khan for her work on the generalization of the auditing framework.

REFERENCES

[1] Ifeoma Ajunwa. 2021. An Auditing Imperative for Automated Hiring Systems. *Harvard Journal of Law & Technology* 34, 2 (2021), 80. https://doi.org/10.2139/ssrn.3437631

[2] Luisa Araujo, Andrea Saltelli, and Sylke V. Schnepf. 2017. Do PISA data justify PISA-based education policy? *International Journal of Comparative Education and Development* 19, 1 (Jan. 2017), 20–34. https://doi.org/10.1108/IJCED-12-2016-0023 Publisher: Emerald Publishing Limited.

[3] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction* 5 (2021). Issue CSCW1. https://doi.org/10.1145/3449148

[4] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 671, 104 (2016). https://doi.org/10.2139/ssrn.2477899

[5] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 1 (1995), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1995.tb02031.x

[6] Miranda Bogen and Aaron Rieke. 2018. *Help Wanted: An Exploration of Hiring Algorithms, Equity and Bias.* Technical Report. Upturn. 75 pages. https://www.upturn.org/static/reports/2018/hiring-algorithms/files/Upturn%20--%20Help%20Wanted%20-%20An%20Exploration%20of%20Hiring%20Algorithms,%20Equity%20and%20Bias.pdf

[7] Shea Brown, Jovana Davidovic, and Ali Hasan. 2021. The algorithm audit: Scoring the algorithms that score us. *Big Data & Society* 8, 1 (Jan. 2021), 2053951720983865. https://doi.org/10.1177/2053951720983865 Publisher: SAGE Publications Ltd.

[8] Jean Cardinet, Yvan Tourneur, and Linda Allal. 1976. The Symmetry of Generalizability Theory: Applications to Educational Measurement. *Journal of Educational Measurement* 13, 2 (1976), 119–135. https://www.jstor.org/stable/1434233 Publisher: National Council on Measurement in Education, Wiley.

[9] Edward Carmines and Richard Zeller. 1979. *Reliability and Validity Assessment.* SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States of America. https://doi.org/10.4135/9781412985642

[10] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 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.* ACM, Montreal QC Canada, 1–14. https://doi.org/10.1145/3173574.3174225

[11] "The 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.

[12] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (Oct. 2018). https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[13] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (April 2015), 92–112. https://doi.org/10.1515/popets-2015-0007

[14] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* ACM, Atlanta GA USA, 120–128. https://doi.org/10.1145/3287560.3287572

[15] Merve Emre. 2018. *The Personality Brokers: The Strange History of Myers-Briggs and the Birth of Personality Testing* (first ed.). Doubleday, New York.

[16] Equal Employment Opportunity Commission (EEOC), Civil Service Commission, Department of Labor, and Department of Justice. 1978. Uniform guidelines on employee selection procedures. *Federal Register* (Aug. 1978). https://www.ojp.gov/pdffiles1/Digitization/58846NCJRS.pdf

[17] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *SSRN Electronic Journal* (2020). https://doi.org/10.2139/ssrn.3518482

[18] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems* 14, 3 (July 1996), 330–347. https://doi.org/10.1145/230538.230561

[19] Alessandro Galli, Mario Giampietro, Steve Goldfinger, Elias Lazarus, David Lin, Andrea Saltelli, Mathis Wackernagel, and Felix Müller. 2016. Questioning the Ecological Footprint. *Ecological Indicators* 69 (Oct. 2016), 224–232. https://doi.org/10.1016/j.ecolind.2016.04.014

[20] Leonard D Goodstein and Richard I Lanyon. 1999. Applications of Personality Assessment to the Workplace: A Review. *Journal of Business and Psychology* 13, 3 (1999), 32.

[21] Robert M. Guion and Richard F. Gottier. 1965. Validity Of Personality Measures In Personnel Selection. *Personnel Psychology* 18, 2 (1965), 135–164. https://doi.org/10.1111/j.1744-6570.1965.tb00273.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1744-6570.1965.tb00273.x.

[22] Richard A. Guzzo, Alexis A. Fink, Eden King, Scott Tonidandel, and Ronald S. Landis. 2015. Big Data Recommendations for Industrial-Organizational Psychology. *Industrial and Organizational Psychology* 8, 4 (Dec. 2015), 491–508. https://doi.org/10.1017/iop.2015.40 Num Pages: 18 Place: Bowling Green, United Kingdom Publisher: Cambridge University Press.

[23] Thilo Hagendorff. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines* 30, 1 (March 2020), 99–120. https://doi.org/10.1007/s11023-020-09517-8

[24] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 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.* ACM, Portland Oregon USA, 1914–1933. https://doi.org/10.1145/2998181.2998327

[25] Leaetta M. Hough, Newell K. Eaton, Marvin D. Dunnette, John D. Kamp, and Rodney A. McCloy. 1990. Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology* 75, 5 (1990), 581–595. https://doi.org/10.1037/0021-9010.75.5.581 Place: US Publisher: American Psychological Association.

[26] Gregory M. Hurtz and John J. Donovan. 2000. Personality and job performance: The Big Five revisited. *Journal of Applied Psychology* 85, 6 (2000), 869–879. https://doi.org/10.1037/0021-9010.85.6.869 Place: US Publisher: American Psychological Association.

[27] Aislinn Kelly-Lyth. 2020. *Challenging Biased Hiring Algorithms.* SSRN Scholarly Paper ID 3744248. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=3744248

[28] Pauline T Kim. 2017. Data-Driven Discrimination at Work. *William & Mary Law* 58 (2017), 81.

[29] Adriano Koshiyama, Emre Kazim, Philip Treleaven, Pete Rai, Lukasz Szpruch, Giles Pavey, Ghazi Ahamat, Franziska Leutner, Randy Goebel, Andrew Knight, Janet Adams, Christina Hitrova, Jeremy Barnett, Parashkev Nachev, David Barber, Tomas Chamorro-Premuzic, Konstantin Klemmer, Miro Gregorovic, Shakeel Khan, and Elizabeth Lomas. 2021. *Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms.* Technical Report. https://www.ssrn.com/abstract=3778998

[30] Max Langenkamp, Allan Costa, and Chris Cheung. 2020. Hiring Fairly in the Age of Algorithms. *arXiv:2004.07132 [cs]* (April 2020). http://arxiv.org/abs/2004.07132 arXiv: 2004.07132.

[31] Samuele Lo Piano and Marguerite Robinson. 2019. Nutrition and public health economic evaluations under the lenses of post normal science | Elsevier Enhanced Reader. *Futures* 112 (2019). https://doi.org/10.1016/j.futures.2019.06.008

[32] Kira Lussier. 2018. Temperamental Workers: Psychology, Business, and the Humm-Wadsworth Temperament Scale in Interwar America. *History of Psychology* 21, 2 (2018), 79. https://doi.org/10.1037/hop0000081 Publisher: US: American Psychological Association.

[33] Dori Meinert. 2015. What Do Personality Tests Really Reveal? https://www.shrm.org/hr-today/news/hr-magazine/pages/0615-personality-tests.aspx

[34] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. https://papers.ssrn.com/abstract=3736261

[35] Frederick P. Morgeson, Michael A. Campion, Robert L. Dipboye, John R. Hollenbeck, Kevin Murphy, and Neal Schmitt. 2007. Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology* 60, 3 (2007), 683–729. https://doi.org/10.1111/j.1744-6570.2007.00089.x Place: United Kingdom Publisher: Blackwell Publishing.

[36] James R. Morrow and Allen w. Jackson. 1993. How "Significant" is Your Reliability? *Research Quarterly for Exercise and Sport* 64, 3 (Sept. 1993), 352–355. https://doi.org/10.1080/02701367.1993.10608821

[37] Ralph O. Mueller and Thomas R. Knapp. 2018. Reliability and Validity. In *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (2 ed.). Routledge. Num Pages: 5.

[38] Jakob Mökander, Jessica Morley, Mariarosaria Taddeo, and Luciano Floridi. 2021. Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. *Science and Engineering Ethics* 27, 4 (Aug. 2021), 44. https://doi.org/10.1007/s11948-021-00319-4

[39] Jum C. Nunnally and Ira H. Bernstein. 1994. *Psychometric Theory* (3 ed.). McGraw-Hill, Inc.

[40] Luis Oala, Jana Fehr, Luca Gilli, Pradeep Balachandran, Alixandro Werneck Leite, Saul Calderon-Ramirez, Danny Xie Li, Gabriel Nobis, Erick Alejandro Muñoz Alvarado, Giovanna Jaramillo-Gutierrez, Christian Matek, Arun Shroff, Ferath Kherif, Bruno Sanguinetti, and Thomas Wiegand. 2020. ML4H Auditing: From Paper to Practice. In *Machine Learning for Health.* PMLR, 280–317. https://proceedings.mlr.press/v136/oala20a.html ISSN: 2640-3498.

[41] O'Neil Risk Consulting and Algorithmic Auditing (ORCAA). 2020. *Description of Algorithmic Audit: Pre-built Assessments.* Technical Report. https://techinquiry.org/HireVue-ORCAA.pdf

[42] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *Proceedings of the*

2020 Conference on Fairness, Accountability, and Transparency. ACM, Barcelona Spain, 469–481. https://doi.org/10.1145/3351095.3372828

[43] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. ACM, Barcelona, Spain, 12. https://doi.org/10.1145/3351095.3372873

[44] Saman Razavi, Anthony Jakeman, Andrea Saltelli, Clémentine Prieur, Bertrand Iooss, Emanuele Borgonovo, Elmar Plischke, Samuele Lo Piano, Takuya Iwanaga, William Becker, Stefano Tarantola, Joseph H. A. Guillaume, John Jakeman, Hoshin Gupta, Nicola Melillo, Giovanni Rabitti, Vincent Chabridon, Qingyun Duan, Xifu Sun, Stefán Smith, Razi Sheikholeslami, Nasim Hosseini, Masoud Asadzadeh, Arnald Puy, Sergei Kucherenko, and Holger R. Maier. 2021. The Future of Sensitivity Analysis: An essential discipline for systems modeling and policy support. *Environmental Modelling & Software* 137 (March 2021), 104954. https://doi.org/10.1016/j.envsoft.2020.104954

[45] Alene K. Rhea, Kelsey Markey, Lauren D'Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, Falaah Arif Khan, and Julia Stoyanovich. 2022. An External Stability Audit Framework to Test the Validity of Personality Prediction in AI Hiring. *CoRR* abs/2201.09151 (2022). arXiv:2201.09151 https://arxiv.org/abs/2201.09151

[46] Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–22. https://doi.org/10.1145/3274417

[47] Andrea Saltelli, Gabriele Bammer, Isabelle Bruno, Erica Charters, Monica Di Fiore, Emmanuel Didier, Wendy Nelson Espeland, John Kay, Samuele Lo Piano, Deborah Mayo, Roger Pielke Jr, Tommaso Portaluri, Theodore M. Porter, Arnald Puy, Ismael Rafols, Jerome R. Ravetz, Erik Reinert, Daniel Sarewitz, Philip B. Stark, Andrew Stirling, Jeroen van der Sluijs, and Paolo Vineis. 2020. Five ways to ensure that models serve society: a manifesto. *Nature* 582, 7813 (June 2020), 482–484. https://doi.org/10.1038/d41586-020-01812-9 Bandiera_abtest: a Cg_type: Comment Number: 7813 Publisher: Nature Publishing Group Subject_term: Communication, Policy, Epidemiology.

[48] Andrea Saltelli and Samuele Lo Piano. 2017. Problematic Quantifications: a Critical Appraisal of Scenario Making for a Global 'Sustainable' Food Production. *Food Ethics* 1, 2 (Aug. 2017), 173–179. https://doi.org/10.1007/s41055-017-0020-6

[49] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. In *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*. Seattle, WA, USA, 23.

[50] Hilke Schellmann, Jennifer Strong, Ian, and Siegel. 2021. Hired by an algorithm. https://podcasts.apple.com/us/podcast/hired-by-an-algorithm/id1523584878?i=1000526571833. *In Machines We Trust podcast series, MIT Technology Review* (2021).

[51] Hilke Schellmann, Jennifer Strong, Ian, and Siegel. 2021. Want a job? The AI will see you now. https://podcasts.apple.com/us/podcast/want-a-job-the-ai-will-see-you-now/id1523584878?i=1000528104144. *In Machines We Trust podcast series, MIT Technology Review* (2021).

[52] Frank L. Schmidt, Huy Le, and Remus Ilies. 2003. Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods* 8, 2 (2003), 206. https://doi.org/10.1037/1082-989X.8.2.206 Publisher: US: American Psychological Association.

[53] Neal Schmitt, Richard Z. Gooding, Raymond A. Noe, and Michael Kirsch. 1984. Metaanalyses of Validity Studies Published Between 1964 and 1982 and the Investigation of Study Characteristics. *Personnel Psychology* 37, 3 (1984), 407–422. https://doi.org/10.1111/j.1744-6570.1984.tb00519.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1744-6570.1984.tb00519.x.

[54] 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. https://doi.org/10.26356/masos

[55] Wesley A. Scroggins, Steven L. Thomas, and Jerry A. Morris. 2008. Psychological Testing in Personnel Selection, Part I: A Century of Psychological Testing. *Public Personnel Management* 37, 1 (March 2008), 99–109. https://doi.org/10.1177/009102600803700107 Publisher: SAGE Publications Inc.

[56] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 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 (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 166–172. https://doi.org/10.1145/3375627.3375812

[57] Ben Shneiderman. 2020. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems* 10, 4 (Dec. 2020), 1–31. https://doi.org/10.1145/3419764

[58] Mona Sloane. 2021. The Algorithmic Auditing Trap. https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d

[59] Mona Sloane, Emanuel Moss, and Rumman Chowdhury. 2022. A Silicon Valley Love Triangle: Hiring Algorithms, Pseudo-Science, and the Quest for Auditability. *Patterns* 3, 2 (2022). https://doi.org/10.1016/j.patter.2021.100425

[60] Luke Stark and Jevan Hutson. 2021. Physiognomic Artificial Intelligence. *SSRN Electronic Journal* (2021). https://doi.org/10.2139/ssrn.3927300

[61] Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK. 2020. *Auditing machine learning algorithms*. Technical Report. 57 pages. https://www.auditingalgorithms.net/

[62] Tom Sühr, Sophie Hilgard, and Himabindu Lakkaraju. 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*. ACM, Virtual Event USA, 989–999. https://doi.org/10.1145/3461702.3462602

[63] Peter Turney. 1995. Technical note: Bias and the quantification of stability. *Machine Learning* 20, 1-2 (1995), 23–33. https://doi.org/10.1007/BF00993473

[64] Tyler J VanderWeele and Maya B Mathur. 2019. SOME DESIRABLE PROPERTIES OF THE BONFERRONI CORRECTION: IS THE BONFERRONI CORRECTION REALLY SO BAD? *American Journal of Epidemiology* 188, 3 (March 2019), 617–618. https://doi.org/10.1093/aje/kwy250

[65] Giridhari Venkatadri, Athanasios Andreou, Yabing Liu, Alan Mislove, Krishna P. Gummadi, Patrick Loiseau, and Oana Goga. 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)*. 89–107. https://doi.org/10.1109/SP.2018.00014 ISSN: 2375-1207.

[66] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 666–677. https://doi.org/10.1145/3442188.3445928

[67] Songkai Xue, Mikhail Yurochkin, and Yuekai Sun. 2020. Auditing ML Models for Individual Bias and Unfairness. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 4552–4562. https://proceedings.mlr.press/v108/xue20a.html ISSN: 2640-3498.

A ADDITIONAL AUDIT DETAILS

A.1 Treatments for each facet

Details of score-generating model calls to generate treatments for each facet, discussed in Section 4.3, are presented in Table 3 for `Humantic AI` and in Table 4 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 executed on this input. We also list the number of inputs submitted ("# In") and the number of profiles produces ("# Out"). 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 the system did not produce personality profiled for 11 of them, and returned errors saying the Twitter profiles were "thin."

A.2 Stability metrics

We used the following metrics to assess facet-specific stability.

- **Rank-order stability.** Because DiSC scores were discontinuous in `Crystal`, we use Spearman's rank correlation rather than Pearson's correlation coefficient to quantify rank-order stability. Rank-order stability results are presented in Tables 5, 6, and 7.
- **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 to be normally distributed. Locational stability results can be found in Tables 8, 9, and 10. 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 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).

A.3 Additional results

A.3.1 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).

A.3.2 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 $p = 7x10^{-6}$) and Extraversion (LinkedIn median 6.44, resume median 5.75; Wilcoxon $p = 6.9x10^{-5}$) , and significant locational differences under Benjamini-Hochberg in Influence (LinkedIn meidan 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 $p = 2.8x10^{-4}$). Although there was no significant locational instability for Agreeableness overall, for non-native English speakers, the median Agreeableness score on resumes (5.99) was significantly different from the median score on LinkedIn (5.63) under Benjamini-Hochberg ($p = 6.1 \times 10^{-3}$).

Comparing `Humantic AI` scores from Twitter to those from original resumes, we find correlations ranging from -0.521 (Dominance)

Table 3: Details of `Humantic AI` runs (i.e., sets of score-generating calls to `Humantic AI` models).

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

Table 4: Details of `Crystal` runs (i.e., sets of score-generating calls to `Crystal` models).

Input Type	Run ID	Run Dates	# In	# Out
Raw Text Resume	CRr1	03/31/21 - 04/02/21	89	89
Raw Text Resume	CRr2	05/01/21 - 05/03/21	89	89
Raw Text Resume	CRr3	05/01/21 - 05/03/21	89	89
PDF Resume	CRp1	11/23/20 - 01/14/21	89	89
LinkedIn	CL1	11/23/20 - 01/14/21	92	91
LinkedIn	CL2	09/13/21 - 09/16/21	89	89

Figure 7: Normalized L1 distances between `Humantic AI` DiSC and Big Five scores produced from identical resumes scored at different points in time.

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 participants' 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.

See Tables 6, 7, 9, and 10 for complete results for `Humantic AI`.

A.3.3 Algorithm time. Figure 7 shows that low sub-group correlations are due to 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 $p = 7.1 \times 10^{-3}$).

A.3.4 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.

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 6, 7, 9, and 10.

Table 5: Rank-order stability of `Crystal` DiSC scores, as measured by Spearman's rank correlations. Reliabilities below 0.90 highlighted in yellow; those between 0.90 and 0.95 highlighted in lighter yellow. Results are discussed in Sections 4.5.4, 4.5.5, 4.5.6, 4.5.7, and 4.5.8.

Facet	Input Versions	N	Dominance	Influence	Steadiness	Conscientiousness
File Format (Resume, raw text vs. PDF)	CRr1 vs. CRp1	89	0.8225	0.8260	0.9184	0.9114
Source Context (Resume vs. LinkedIn)	CRp1 vs. CL1	86	0.2335	0.5258	0.5103	0.3585
Immediate Rep. (Resume)	CRr2 vs. CRr3	89	1.0000	1.0000	1.0000	1.0000
Algorithm-Time (Resume)	CRr1 vs. CRr2	89	1.0000	1.0000	1.0000	1.0000
Participant-Time (LinkedIn)	CL1 vs. CL2	89	0.5314	0.7062	0.8676	0.7811

Table 6: Rank-order stability of `Humantic AI` DiSC scores, as measured by Spearman's rank correlations. Reliabilities below 0.90 highlighted in yellow. Results are discussed in Sections 4.5.4, 4.5.5, 4.5.6, 4.5.7, and 4.5.8.

Facet	Input Versions	N	Dominance	Influence	Steadiness	Calculativeness
File Format (Resume, de-id vs. DOCX)	HRi1 vs. HRd1	89	0.9956	0.9924	0.9978	0.9959
URL Embedding (Resume)	HRu1 vs. HRi1	86	0.3570	0.6253	0.5480	0.6878
URL Embedding (Resume vs. LinkedIn)	HRu1 vs. HL1	83	0.1555	0.3382	0.6074	0.4701
Source Context (Resume vs. LinkedIn)	HRi1 vs. HL1	84	0.0903	0.2553	0.3941	0.3331
Source Context (Resume vs. LinkedIn)	HRo1 vs. HL1	84	0.1775	0.4016	0.6939	0.6249
Source Context (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. (Resume)	HRi2 vs. HRi3	89	0.9999	1.0000	1.0000	1.0000
Algorithm-Time (Resume)	HRi1 vs. HRi2	89	0.9726	0.9948	0.9925	0.9980
Participant-Time (LinkedIn)	HL1 vs. HL2	88	0.2248	0.4186	0.6597	0.5827
Participant-Time (Twitter)	HT1 vs. HT2	21	1.0000	1.0000	1.0000	1.0000

Table 7: Rank-order stability of `Humantic AI` Big Five scores, as measured by Spearman's rank correlations. Reliabilities below 0.90 highlighted in yellow. Results are discussed in Sections 4.5.4, 4.5.5, 4.5.6, 4.5.7, and 4.5.8.

Facet	Input Versions	N	Openness	Conscientiousness	Extraversion	Agreeableness	Emotional Stability
File Format (Resume, de-id vs. DOCX)	HRi1 vs. HRd1	89	0.9891	0.9936	0.9939	0.9927	0.9816
URL Embedding (Resume)	HRu1 vs. HRi1	86	0.3988	0.3845	0.0772	0.4190	0.4040
URL Embedding (Resume vs. LinkedIn)	HRu1 vs. HL1	83	0.6381	0.5470	0.5786	0.6839	0.7018
Source Context (Resume vs. LinkedIn)	HRi1 vs. HL1	84	0.2180	0.1558	0.1198	0.2020	0.2186
Source Context (Resume vs. LinkedIn)	HRo1 vs. HL1	84	0.5985	0.7124	0.5827	0.6136	0.5990
Source Context (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. (Resume)	HRi2 vs. HRi3	89	1.0000	1.0000	1.0000	0.9999	1.0000
Algorithm-Time (Resume)	HRi1 vs. HRi2	89	0.9954	0.9969	0.9618	0.9921	0.9854
Participant-Time (LinkedIn)	HL1 vs. HL2	88	0.6879	0.6928	0.7301	0.7518	0.7678
Participant-Time (Twitter)	HT1 vs. HT2	21	1.0000	1.0000	1.0000	1.0000	1.0000

Table 8: Locational stability of `Crystal` DiSC scores, as measured by two-tailed Wilcoxon signed-rank test p-values. The absence of yellow highlighting indicates that all values are above both the Benjamini-Hochberg and Bonferroni-corrected thresholds based on α of 0.05. "N/A" values reflect experiments where there was no change across the facet, indicating perfect stability. Results are discussed in Sections 4.5.4, 4.5.5, 4.5.6, 4.5.7, and 4.5.8.

Facet	Input Versions	N	Dominance	Influence	Steadiness	Conscientiousness
File Format (Resume, raw text vs. PDF)	CRr1 vs. CRp1	89	0.5026	0.4208	0.0173	0.0370
Source Context (Resume vs. LinkedIn)	CRp1 vs.CL1	86	0.4190	0.0012	0.7010	0.8421
Immediate Rep. (Resume)	CRr2 vs. CRr3	89	N/A	N/A	N/A	N/A
Algorithm-Time (Resume)	CRr1 vs. CRr2	89	N/A	N/A	N/A	N/A
Participant-Time (LinkedIn)	CL1 vs. CL2	89	0.7299	0.6518	0.3305	0.2870

Table 9: Significance in locational instability of `Humantic AI` DiSC scores, as measured by two-tailed Wilcoxon signed-rank test p-values. Yellow highlighting indicates value below Bonferroni-corrected threshold based on α of 0.05. Lighter yellow indicates p-value below Benjamini-Hochberg corrected threshold and above Bonferroni-corrected threshold. "N/A" values reflect experiments where there was no change across the facet. Results are discussed in Sections 4.5.4, 4.5.5, 4.5.6, 4.5.7, and 4.5.8.

Facet	Input Versions	N	Dominance	Influence	Steadiness	Calculativeness
File Format (Resume, de-id vs. DOCX)	HRi1 vs. HRd1	89	0.2510	0.2940	0.4574	0.2539
URL Embedding (Resume)	HRu1 vs. HRi1	86	0.0000	0.3194	0.0005	0.0047
URL Embedding (Resume vs. LinkedIn)	HRu1 vs. HL1	83	0.0066	0.1825	0.5324	0.1213
Source Context (Resume vs. LinkedIn)	HRi1 vs. HL1	84	0.0000	0.0580	0.0013	0.3259
Source Context (Resume vs. LinkedIn)	HRo1 vs. HL1	84	0.0000	0.0050	0.2299	0.5911
Source Context (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. (Resume)	HRi2 vs. HRi3	89	0.3173	0.3173	N/A	N/A
Algorithm-Time (Resume)	HRi1 vs. HRi2	89	0.1416	0.5971	0.5690	0.0307
Participant-Time (LinkedIn)	HL1 vs. HL2	88	0.0709	0.0800	0.3457	0.2969
Participant-Time (Twitter)	HT1 vs. HT2	21	N/A	N/A	N/A	N/A

Table 10: Significance in locational instability of `Humantic AI` Big Five scores, as measured by two-tailed Wilcoxon signed-rank test p-values. Yellow highlighting indicates value below Bonferroni-corrected threshold based on α of 0.05. Lighter yellow indicates p-value below Benjamini-Hochberg corrected threshold and above Bonferroni-corrected threshold. "N/A" values reflect experiments where there was no change across the facet. Results are discussed in Sections 4.5.4, 4.5.5, 4.5.6, 4.5.7, and 4.5.8.

Facet	Input Versions	N	Openness	Conscien-tiousness	Extraversion	Agreeableness	Emotional Stability
File Format (Resume, de-id vs. DOCX)	HRi1 vs. HRd1	89	0.7193	0.9248	0.5306	0.3003	0.9771
URL Embedding (Resume)	HRu1 vs. HRi1	86	0.0025	0.0000	0.0000	0.0002	0.2214
URL Embedding (Resume vs. LinkedIn)	HRu1 vs. HL1	83	0.7352	0.0000	0.3603	0.0068	0.7167
Source Context (Resume vs. LinkedIn)	HRi1 vs. HL1	84	0.0077	0.3997	0.0000	0.1730	0.6718
Source Context (Resume vs. LinkedIn)	HRo1 vs. HL1	84	0.5300	0.0003	0.0001	0.0221	0.4553
Source Context (Resume vs. Twitter)	HRo1 vs. HT1	20	0.0121	0.0826	0.8983	0.0020	0.0010
Source Context (LinkedIn vs. Twitter)	HL1 vs. HT1	18	0.0023	0.0047	0.0007	0.0047	0.0007
Immediate Rep. (Resume)	HRi2 vs. HRi3	89	0.1797	0.3173	0.3173	0.6547	0.6547
Algorithm-Time (Resume)	HRi1 vs. HRi2	89	0.0071	0.5314	0.2540	0.0516	0.2424
Participant-Time (LinkedIn)	HL1 vs. HL2	88	0.6487	0.0000	0.9615	0.0072	0.6011
Participant-Time (Twitter)	HT1 vs. HT2	21	N/A	N/A	N/A	N/A	N/A