ORCAA

Description of Algorithmic Audit: Pre-built Assessments

December 15, 2020

This report describes ORCAA's algorithmic audit of HireVue's pre-built assessments for early career and campus hires. We summarize the audit process, recap key findings, and describe the steps HireVue is taking to address these findings.

About ORCAA

O'Neil Risk Consulting and Algorithmic Auditing (ORCAA) is a consultancy that helps organizations identify and manage risks arising from the use of predictive models, AI, and related technologies. Our algorithmic audits consider what it means for a model/AI to succeed and how it could fail, with a focus on ethical dimensions including fairness, bias, and discrimination. The audit process engages internal and external stakeholders directly to elicit, evaluate, and address concerns related to a model/AI being deployed in a specific context. We have completed audits in industries including hiring, insurance, and housing/hospitality, as well as with municipal and public agencies. ORCAA is led by Cathy O'Neil, author of *Weapons of Math Destruction*.

Recap of the audit

In April and May, 2020, ORCAA performed an algorithmic audit for HireVue. The audit focused on fairness and bias issues around a specific use case: pre-built assessments used in hiring early career candidates, including from college campuses. These assessments incorporate algorithms that analyze candidates' responses to interview questions recorded via webcam, and their performance on psychometric games, to generate competency scores in eight areas: Communication, Team Orientation, Problem Solving, Willingness to Learn, Adaptability, Dependability, Drive for Results & Initiative, and Cognitive Ability. In the use case audited, the competency scores are then used to rank the candidates, with the highest-ranked progressing to an in-person interview, the lowest-ranked being rejected, and the remainder referred to a human reviewer.

This was not a comprehensive audit of HireVue's use of algorithms; our findings are particular to the use case we considered. We note three qualifications along these lines. First, in terms of HireVue's offerings, pre-built assessments are distinct from

custom assessments, which are designed around job-related outcomes specified by the client, e.g., sales levels, tenure, or performance review scores. Instead of scoring generic competency areas like Adaptability, a custom assessment might predict what a candidate's job performance would be, were that candidate hired. This audit did not cover custom assessments. Second, the use case we audited is not necessarily common or representative of HireVue's business overall. ORCAA encouraged HireVue to choose a "challenging" use case for the audit – one that would prompt hard fairness questions – since we have found working through complex examples produces more valuable lessons. Finally, the audit considered validity as it relates to fairness and bias concerns. Specifically we investigated whether the competency score models used in pre-built assessments accurately predict the competency scores candidates would have been given by trained human reviewers, or would have achieved in psychometric tests, while upholding fair hiring standards (they do).

Process

The algorithmic audit can be seen as an extended conversation that ORCAA helped to facilitate and document. The conversation comprised four steps:

1. **Preparation** // Defining the use case and reviewing background documentation provided by HireVue.
2. **Stakeholder discussions** // Interviewing stakeholders of the algorithm to elicit their concerns. Essentially this means asking, "How could this fail, and what would that mean for you?" Stakeholders included teams within HireVue (e.g., data science, I/O Psychology, legal, product) and external stakeholders (Integrate Advisors, representing neuro-atypical candidates; Jopwell and re:work, representing minority candidates; a HireVue client, representing the customer perspective; an AI fairness researcher).
3. **The Ethical Matrix** // Mapping stakeholders and their concerns onto a grid[1], which ORCAA and HireVue discussed to prioritize the most pressing concerns. The key question was, "Which concerns, if realized, would be an existential threat to the algorithm working, or a 'dealbreaker' for some stakeholder?"
4. **Planning remediation steps** // Coming up with ways to investigate, validate, and address the priority concerns identified. ORCAA met with domain experts at HireVue about internal processes and reviewed documentation on their processes for preparation, validation, and testing of their algorithms. The result of this phase was a set of task lists, each corresponding to a priority concern, that name the specific steps HireVue plans to take.

Results of the Audit: Areas of concern and remediation steps

The main findings of the audit were:

[1] The Ethical Matrix and overall audit process are described in more detail in "The Ethical Matrix Process," available upon request.

- The assessments work as advertised with regard to fairness and bias issues; ORCAA did not find any operational risks with respect to clients using them.
- Score gaps between demographic groups is a key fairness risk. Through deliberate feature selection and engineering as well as bias mitigation, HireVue ensures that every assessment complies with the major legal fair-hiring standard with regard to race, ethnicity, gender, and age. Other protected characteristics (e.g. religion, disability) are considered when possible, which varies depending on customer data availability.
- Making assessments work for everybody – no matter their level of career experience, disability status, tech-savviness, etc. – is an ongoing project with important fairness implications. Relatedly, those who opt out of HireVue assessments could represent a fairness concern.
- Some candidates don't understand what the assessment is "looking for" or exactly how it fits in the hiring process. These may be generic (vs. HireVue-specific) job-seeking concerns; nonetheless, providing more information could help these candidates.

Below we describe the key areas of potential concern stakeholders raised and the steps HireVue has identified to address them.

Concern + Remediation Steps: Bias in model scores

Stakeholders indicated that it would be a major fairness risk if the competency scores given by the algorithm had systematic differences across demographic groups. This would also be a legal risk if there were score gaps by gender, age, race, or ethnicity, four legally protected classes that are often the focus of fair hiring cases. The key standard here is the EEOC's Uniform Guidelines on Employee Selection Procedures, which define adverse impact in this context as "employment practices that appear neutral but have a discriminatory or negative effect on a protected group." It occurs when a selection, hiring, or promotion decision results in substantial differences or unequal outcomes for a protected class or group. The "⅘ rule" is used as a practical test for adverse impact: it says that a group experienced an adverse impact if the selection rate for that group was less than ⅘ that of the group with the highest selection rate. Other legally protected classes such as disability status, religion, and national origin were also mentioned in this vein. Since data on these other classes are often not collected during job application processes, and therefore are not available to HireVue, and are also not easily inferred, the action items around them are more exploratory.

HireVue addresses this issue mainly through the model development and QA process – in particular, reviewing the scores given by the statistical models and changing the models as necessary. Before the audit, HireVue already had such measures in place to ensure assessments are consistent with EEOC's Uniform Guidelines with regard to gender, age, race, and ethnicity groups. They plan to build on this with research into

the relationship between accents and competency scores assessed by video (accent does not play a role in competencies assessed by psychometric games).

Step	Explanation	Status
Testing for adverse impact and group differences in scores	*Competency scores are reviewed before and during deployment of each model to ensure that the assessment satisfies standards set forth in the EEOC Uniform Guidelines and I/O professional testing standards with regard to gender, age, race, and ethnicity. This includes ensuring that there is not significant evidence of adverse impact on any particular group in which candidates are disqualified vs. allowed to proceed to the next step in the hiring process. Two kinds of evidence are considered: (1) measures of practical significance[2], which speak to the magnitude of the difference between groups; and (2) measures of statistical significance[3], which speak to the likelihood that the difference between groups is due to random chance. If an assessment finds evidence of both practically and statistically significant adverse impact, HireVue will not release it. If results are inconclusive or partial evidence is found, HireVue will recommend that the assessment not be used to disqualify candidates, and will work to monitor, mitigate, or gather more data. When choosing between models that do not indicate adverse impact, other metrics are considered such as Cohen's D (differences across groups in mean scores should be small) and "Ranking AI", a novel metric that summarizes whether the model tends to rank members of any particular demographic group above others in a significant way (it shouldn't).*	Already done pre-audit
Removing proxies	*Each variable in a predictive model is tested for its correlation with race, gender, and age. Any variables that are highly correlated and are not very predictive of the job-related outcome are left out of the model.*	Already done pre-audit
Balanced training data	*Training data for competency models are intentionally sampled to make sure groups have good representation in terms of age, race, gender, country of residence, and job role type.*	Already done pre-audit
Study of accents	*HireVue is reviewing its own data to assess whether there are score differences according to accent. They have already compared native English speakers vs non-native speakers; they have not yet looked at regional accents among English speakers.*	In progress
	Working with a transcription vendor to understand how accents in spoken video clips might influence competency scores.	In progress
Explaining bias testing + fixes	*HireVue is currently writing a paper for submission to an I/O Psychology journal that describes their adverse impact remediation process.*	In progress

[2] The primary measure used is the Adverse Impact Ratio (AIR), defined as the ratio of passing rates between the lowest- and highest-performing groups within a given protected class. For instance, for gender, if 75% of men pass and 70% of women pass, then AIR=0.70/0.75=0.93. If AIR is above 0.8 with a robust sample (50+ individuals in each group and 400+ overall), then there is no evidence of practically significant adverse impact. If AIR is below 0.8 with a small sample, or slightly/imprecisely below 0.8 with reassuring evidence from other metrics (Odds Ratio 0.7-1.4 and Cohen's h<|0.2|), then there is inconclusive evidence of practically significant adverse impact. Otherwise there is evidence of practically significant adverse impact.

[3] The measure used here is a z-test for difference in proportions, where the proportions are passing rates of the lowest- and highest-performing groups within a protected class. HireVue uses a significance level of 5% in these tests.

Concern + Remediation Steps: Ensuring assessments work for all job seekers

Stakeholders worried that some people could be disadvantaged by the content or format of a pre-built assessment. For instance, less tech-savvy candidates might struggle with the app-style games, or candidates with speech disorders might find video recording disconcerting. Such a disadvantage could result in score gaps (as just discussed) but not necessarily; concerns in this area focused on differences in how candidates experience the assessment, and whether their applications could be scored at all. Broadly, these concerns have to do with accessibility.

Related to this area of concern is a technique called "thresholding" – pre-screening candidates' responses to video interview questions to confirm they contain enough content to generate meaningful competency scores via algorithm. If not (e.g., if the candidate's voice is inaudible, or if the clip is very short) the application will be "thresholded" and set aside to be scored by a human reviewer instead. HireVue introduced thresholding previously to address the concern that some groups could be disadvantaged by algorithmic scoring of low-content answers. About 5% of all videos get thresholded; analysis prompted by the audit found there were differences across ethnicities in the rate of thresholding, which was addressed as described in "Dealing with short answers" below.

Step	Explanation	Status
Phasing out visual model features	*Historically some models include features derived from visual data, e.g. the way a candidate's face moves could contribute to their score. There were concerns this might make candidates uncomfortable in general; and that candidates with head or face coverings would be disproportionately flagged for human review, and it is unclear whether that would help or hurt their chances. These features are no longer being built into new models, and are being removed from existing models on a rolling basis as models come up for annual review.*	Began before audit; phase-out via annual review is ongoing
Dealing with short answers	*HireVue discovered a key driver of group differences in thresholding rates was very short answers (e.g. "I don't know.") Such answers were given disproportionately by minority candidates, and they were being thresholded (flagged for human review) instead of scored. After confirming that short answers were adequately represented in the training data to support accurate scoring, now these are being scored via algorithm, which has closed the gap in thresholding rates.*	Done since audit
	HireVue is considering introducing dynamic follow-up questions (e.g., "Could you say more?") in the event of short answers.	Planned
Advocacy group partner-ships	*HireVue works with advocacy groups for candidates of color and neuro-atypical candidates, gathering their input and feedback to improve the accessibility of assessments.*	Done pre-audit; ongoing
	Specifically, Integrate Advisors reviewed video interview questions, flagging words or phrases that might be misunderstood or pose other problems for	Done pre-audit

	neuro-atypical candidates. These questions could then be reworded.	
Improving synonym robustness	*HireVue is upgrading the linguistic components of its models so scores are less sensitive to (arbitrarily) specific words. This is a complex issue, since some words are very important. For instance, perhaps saying 'soda' vs 'pop' should not affect scores, but saying 'client' vs 'customer' should.*	Done since audit

Concern + Remediation Steps: Candidates opting out

Stakeholders had concerns about candidates that can't or don't want to take a HireVue assessment. Do these candidates disproportionately come from certain demographic groups? What alternative(s) are they offered if they still want to apply for the job? These questions are relevant to the fairness of the hiring process even if HireVue's assessment is completely equitable to all candidates who take it.

Step	**Explanation**	**Status**
Improved signposting	*Candidates will be told in advance exactly what capabilities they need to complete the assessment activities (e.g. "To complete this game, you will need to tap moving objects on your touchscreen.")*	In progress
	HireVue revised the instructions candidates are given to set up their assessment. Now candidates are explicitly encouraged to contact the hirer's Diversity & Inclusion team to discuss their options before starting the assessment. That team can decide and coordinate accommodations as they deem appropriate.	Planned
Analysis and encourage-ment	*By studying its historical data, HireVue found suggestive evidence that some groups of candidates are more likely to abandon an assessment after starting. (Note that other groups might be more likely to dropout* before *starting; HireVue's data cannot say.) Next they plan to test in-product encouragement to keep candidates engaged and get them to finish the assessment.*	In progress
Direct survey	*To further explore differential attrition, HireVue will conduct an original survey of job-seekers to see whether there are group differences in willingness to complete virtual assessments as part of the hiring process.*	Planned

Concern + Remediation Steps: Setting candidates up for success

The key questions here were: Do candidates understand how they are being assessed and how to be successful? Are they given enough information to decide whether they can and want to do the assessment? Beyond its current practice, described just below, the following table summarizes additional steps HireVue is taking to address these questions.

Currently, candidates starting an assessment are shown a short video explaining how it works and what to expect. They are also offered a practice question before starting the

video interview portion to get comfortable with the format. This lets them record themselves answering a mock interview question with a time limit. The candidate can then review the video but it is never seen by anyone else. In addition to these in-product features HireVue has published on its website and blog[4] fairly detailed accounts of how its assessments work, and its guiding AI Ethical Principles[5].

Step	Explanation	Status
Clearer instructions	*Before the games portion of the assessment candidates are now shown a video demonstrating gameplay. New in-product messaging explains What's Next in the interview so candidates better understand the experience and know whether or not they are being evaluated by an algorithm*	Completed since the audit
College partner-ships	*A new partnership with college career centers will give students free access to a practice assessment, including video interview questions and a game, without feedback reports. Another upcoming partnership, with Historically Black Colleges and Universities (HBCUs), will give students free accounts that include practice video interview questions and games, with feedback reports.*	Completed since the audit
Expanded practice	*HireVue will add the option to practice one or more games before starting the games portion of the assessment.*	Planned
Giving more info	*To give candidates more information before they decide whether to proceed, HireVue will test messaging around how many other people have already interviewed for the position.*	Planned

In Closing: Context around the Audit

We mention two caveats around the results just presented. First, ORCAA accepted HireVue's assurances regarding the status of remediation steps; we have not independently validated their implementation. Second, stakeholders voiced some minor concerns around data security, e.g., candidate videos could be exposed in a data breach. HireVue has undergone SOC 2 Type 2, ISO 27001, and FedRamp certifications, which it considers to be among the strictest security standards. ORCAA accepted HireVue's assurances on this issue as cybersecurity is not our expertise.

More generally, for the issues of fairness and bias that this audit focused on there are few hard and fast rules. While clear legal standards (e.g. the ⅘ rule) are easier to audit and were included in this audit, much fairness work is a continual improvement process companies navigate – not a checklist provided by regulators or lawmakers. There will be grey areas and ethical dilemmas; it comes with the territory. Thus, "Areas of concern" should not be interpreted as looming risks, or failing to meet some widely-agreed standard. Rather they are the result of engaging with and listening to stakeholders.

Finally, we note the broad context around this work. That HireVue's pre-built assessments are auditable in the first place – based on documented models and

[4] https://www.hirevue.com/blog/creating-ai-driven-pre-employment-assessments
[5] https://www.hirevue.com/why-hirevue/ethical-ai

procedures, against which fidelity can be assessed – is a tractable starting point compared with many hiring processes that rely more on individual human judgments. Indeed, most companies do not undertake an independent algorithmic audit at all. HireVue was not required to do so; it is to their credit that they chose to. In our view, pursuing an audit voluntarily, and acting on the findings, is evidence that HireVue cares about issues of fairness and bias and is doing something about them.