Request for Information (RFI) on Public and Private Sector Uses of Biometric Technologies: Responses

DISCLAIMER: Please note that the RFI public responses received and posted do not represent the views or opinions of the U.S. Government, the Office of Science and Technology Policy (OSTP), or any other Federal agencies or government entities. We bear no responsibility for the accuracy, legality, or content of these responses and the external links included in this document. Additionally, OSTP requested that submissions be limited to 10 pages or less. For submissions that exceeded that length, the posted responses include the components of the response that began before the 10-page limit.
Why was your job application rejected? Bias in Recruitment Algorithms

Humans are biased and so can be the algorithms they develop and the data they use, but what does that mean to you as a job applicant coming out of school or looking to move up to the next step in career ladder or considering a change in roles or industry. What does that mean for society?

In the fast-moving world of technology, AI has particularly expanded into many domains of our personal and business life. Whether you are aware of it or not, algorithmic decision-making systems are now prominently used by companies as well as governments to make decisions on credit worthiness, housing, recruitment, immigration, healthcare, criminal justice system, pricing of goods, welfare eligibility, college admissions – just to name a few. Despite the high stakes and high impact of these decisions on an individual (not to mention the society as a whole), the landscape still greatly lags behind in developing oversight, transparency and accountability measures around these algorithmic systems.¹

Recruitment has always been one of those areas where people always run the risk of being on the receiving end of a biased decision. Given the historical context on discrimination in recruitment, there are a number of legislation actions which provide guidelines and red lines to organizations to make better choices. Title VII of the Civil Rights Act (Civil Rights Act, 1964) for example, prohibits discrimination based on “race, color, religion, sex, or national origin” that would result in disparate treatment or disparate impact. It also puts the liability and legal responsibility on employers to ensure that the tools being used are not creating such results. Moving inside the organizations, in an effort to be more fair, equitable and diverse many companies have taken upon themselves to improve their hiring processes and eliminate as much as possible perceived legacy structural issues through new technology. There is still a long way to go for both the legislation as well as the products and business processes to create more objective and less biased decisions in hiring. Today, companies find AI tools which are used in the full spectrum of hiring process attractive without understanding the potential issues these products might create, or without considering how these results might actually be in complete disagreement with what they want to do with their workforce and future.

¹ Adapted from original article by the author published in Medium, and Montreal AI Ethics Institute blog.

Merve Hickok @ AEthicist.org
The attractiveness of AI-powered recruitment products come from the fact they help companies reach multiple times more candidates than they could reach with the more traditional ways (corporate career websites, referral programs etc.). On the same token, they also make it extremely easy for prospective candidates to submit their CVs to multiple roles at a time with the click of a button. The result is a mutual technology escalation from both the employers and candidates. Once the net is cast, these products help the companies to efficiently process those candidates through the recruitment funnel. The ability to process hundreds of applications in a matter of minutes with an automated system is not only a great benefit in terms of scalability but it also reduces the time to hire and hence potentially the cost to hire (assuming the choices were right) and gives hiring teams more space to develop strategies rather than constantly trying to stay on top of hiring transactions.

One other way these AI recruitment products are marketed is as an alternative to the biased decisions of hiring managers and recruiters and thus to provide a more standard processing of applications. The big issue with this statement is algorithms are not independent of their creators and their biases, nor are they independent of the historical data used to build its models. Algorithms are created by people, about people, for people. In other words, “Algorithms are opinions embedded in code” (Cathy O’Neill, 2017) For long years now, companies have rolled out a number of initiatives to fight the subject and biased decisions involved in hiring decisions. These included training recruiters and hiring managers about unconscious bias so they would be aware of their bias and intentionally and proactively make decisions which are more objective; blinding/hiding certain fields in resumes or applications so hiring managers would not be biased with names, addresses, universities etc; forcing hiring manager and recruiters to have an equal number of male and female candidates in each stage of the recruitment funnel; creating roundtables of hiring committees where candidates are scored against objective criteria and the committee challenges each other on their scores. All these initiatives have some merit in them, but the success varies with the intensity of the effort as well as the culture of the organization trying to make a change. So, when AI systems have potential to reduce bias and reduce cost at the same time, the kneejerk reaction of some companies to jump on wagon without asking too many questions is only too natural.

The AI application in each stage of recruitment may be a recommender system using “collaborative filtering” which makes recommendations based on historic preference of multiple users for items (clicked, liked, rated, etc.), or a “content-based” recommendation by matching for example key
words in your profile or resume. The algorithm might also be a predictive system which uses historical data to find trends or patterns which are then used to predict usually the future, or the likelihood of something happening (for example analyzing the characteristics of applicants who were hired previously and predict your alignment with their success factors, and hence your future success in the job accordingly). It can generate scores or rankings for example for individuals. Alternatively, the AI system might be using a classification algorithm where it maps the input data (in this case candidates) into different categories or clusters. Imagine the recommender systems as Netflix/Pandora where the machine learning algorithm tries to learn your taste and choice by looking at your historical behavior interacting with that app, and it also analyses people whose choices are similar as yours and refines its recommendation; and the predictive systems as your credit scoring.

It is extremely crucial to remember that with AI systems any outcome or prediction is based upon the training data fed into the system. Nature, context and quality of training data for predictive tools can vary, ranging from click patterns, to historical application data, to past hiring decisions, to performance evaluations and productivity measures.

When you add the errors and biased decisions humans made in the past which made up the dataset to the efficiency of the AI systems, you can appreciate how algorithms can magnify the biased decisions. As they are quoting Timothy Wilson (Strangers to Ourselves (2004) in their paper the authors of “Discrimination In The Age Of Algorithms” (Kleinberg & Jens Ludwig & Sendhil Mullainathan & Cass R. Sunstein, 2019) suggest when humans are making decisions, “many choices happen automatically; the influences of choice can be subconscious; and the rationales we produce are constructed after the fact and on the fly”. The researchers than suggest “the black-box nature of the human mind also means that we cannot easily simulate counterfactuals. If hiring managers cannot fully understand why they did what they did, how can even a cooperative manager answer a hypothetical about how he would have proceeded if an applicant had been of a different race or gender” (Kleinberg et al, 2019). The historical record does not help the case for objectivity or fairness with regards to the employers either. In their analysis of trends in discrimination by performing a meta-analysis on 24 field experiments performed between 1990-2015, which included data from more than 54,000 applications across more than 25,000 positions, Quillian and etc al found there were no changes in hiring rates over time for black applicants over the last 25 years. (Quillian, Pager, Hexel, Midtbøen, 2017). In the words of Meredith Whittaker,
co-founder of the AI Now Institute, “AI is not impartial or neutral. In the case of systems meant to automate candidate search and hiring, we need to ask ourselves: What assumptions about worth, ability and potential do these systems reflect and reproduce? Who was at the table when these assumptions were encoded?” (Rosenbaum, 2018)

The following is a step-by-step review of recruitment funnel activities bias can enter the process and result in unintended outcomes.

**TARGETING:**
This is the step when a recruiter tries to cast as wide as a net to the active and passive applicants which would be a strong match to the position that he/she is hiring for. In the digital age, this process has moved from advertising open positions and job descriptions in a company’s corporate website and company profile to publishing it in different career platforms, and general and niche job boards. For any candidate, active or passive, it is crucial the person sees the posting and hence is aware of the opportunity. If you are not aware of the opportunity in the first place, your chances of getting the role is close to nil.

The data collected from your overall online activity provides the platforms with a way to create groups of users with shared attributes (or characteristics, preferences, interests, etc). Today employers have access to the same microtargeting tools advertisers long had on these job boards (like LinkedIn, Glassdoor, ZipRecruiter, Upsider to name just a few of the most known ones). They can select a number of targeting criteria like job seniority, age, gender, degree, etc, and advertise the job opening to candidates in the board’s database.

In 2018, Facebook faced a lawsuit which alleged the social media platform’s practice of allowing job advertisers to consciously target online users by gender, race, and zip code constituted evidence of intentional discrimination (Heater, 2019). Notwithstanding the bias of the recruiter in selecting those criteria and their relevance to job success, the machine learning algorithm in these criteria can collect data on users’ search histories or demographics and use algorithms to predict which individuals companies might want to recruit and only show job postings to those candidates. So as an active or passive candidate if your previous job clicks were significantly more in say junior positions, or in a certain department, the chances you will be targeted for the new job opening are smaller for more senior positions or in different departments. The algorithm also learns from the
recruiter’s behavior and which previous criteria was clicked and used more in previous postings and suggests those to the recruiter. If you are not intentionally making an effort to go over each criterion and verify it, soon your former behavior becomes a personalized default. “It’s part of a cycle: How people perceive things affects the search results, which affect how people perceive things,” Cynthia Matuszek, a computer ethics professor at University of Maryland and co-author of a study on gender bias in Google image search results says (Carpenter, 2015).

Facebook also offers a tool called “lookalike audience” where an employer, might provide Facebook data on its current employees. As Pauline Kim describes it, Facebook takes the source audience, analyzes data about them and identifies other users who have similar profiles, and targets ads to this “lookalike” group to help employers predict which users are most likely to apply for their jobs (Kim, 2018)

On top of all these, the digital marketing platforms like Google or Facebook use their own marketing algorithms to decide which ads are more likely to be clicked by which users within each of their user groups. So just because a recruiter selected ‘all females in Chicago with 10 years of work experience in consulting’ does not mean all those females who fit in that category will see the job posting in their feeds. One experiment by the Carnegie Mellon researchers showed that Google displayed adverts for a career coaching service for “$200k+” executive jobs 1,852 times to the male group and only 318 times to the female group. So in other words, these platforms run their own predictions and further narrow the visibility of a job posting (Vincent, 2015)

**MATCHING & SOURCING:**

If your application made it to the next stage where the candidates are filtered on how much they aligned with the recruiter’s choices, the next set of bias arises from how the algorithm is structured in a way that rank-ordered lists and numerical scores may influence recruiters (Bogen and Rieke, 2018). If a recruiter sees a 95% compatibility vs an 85% compatibility to the job posting, he/she might not even bother to compare the two applications and actually read the totality of the resumes. This issue might snowball further if say the results show only the 10 top ranked applicants per page versus more and the recruiter does not even click to see the rest. On a separate note, when predictions, numerical scores, or rankings are presented as precise and objective, recruiters may give them more weight than they truly warrant, or more deference than a vendor intended (Joachims, Granka, Pan, Hembrooke, and Gay, 2005.) The problem of the algorithm learning from
the recruiter’s previous use of filters for candidate matching (i.e. location, skill, previous company, within x mile radius) is also present in this stage and make future recommendations accordingly.

SCREENING:
So let’s assume your application was one of the ones ranked high in the matching and sourcing platform, and the recruiter clicked your name to process you in to the next stage where you are screened against the company’s preferred criteria. Whether it is through hard-coded questions and filters built into the system, or machine learning algorithms which make decisions, the screening process helps to reduce the number of applications as it goes through your CV / resume and picks up the skills and information (degree, GPA, years of experience, fluency in spoken or technical languages, etc). Whatever the software was able to read (or parse) from your CV, the data points are then matched with the desired points for the specific role. The candidates who have matching points may then again be ranked according to the degree or percentage of match. However, the bigger bias issues in this stage have to do with data out of which the algorithm was created and what kind of a model makes the predictions.

One way to create the datasets by the AI vendors is to scrape data online or buy commercially available datasets – which means a lot of the vendors are using the same data sets. Volume does not mean quality, however. In 2016, Microsoft and Boston University researchers revealed that the Word2Vec (publicly available algorithmic model built on millions of words scraped from online Google News articles, which computer scientists commonly use to analyze word associations) model trained itself on gender stereotypes existing in online news sources (Bolukbasi, 2016). The other finding from the study was these biased word associations were overwhelmingly job related. For example, Man Is to Computer Programmer as Woman is to Homemaker. The data used in training might not have a fair representation in the first place and have embedded bias and imbalances in it, or even if it is perfectly clean it might not be representative of the population you are targeting. In other words, the dataset collected in US might not make a sense if you are a recruiter trying to use this algorithm in Southeast Asia.

Another approach to create the dataset and then the criteria upon which the model is based can be used to look at an organization’s current and past workforce (and/or applicant pool) and determine the success stories and create a model (baseline criteria) based on what “worked” in the past. This is a customized model for the specific employer. However, defining what “worked” or what defines a “successful” employee is also a biased process in itself. What does the client value?
Sales numbers? Cultural fit? Retention? And crucially, what data does the client have? (Raghavan et al, 2020)

“Cultural fit” is a term which is used so frequently we forget it is a subjective measure. It is a better way of saying we will hire people who are like us, or we will not step out of our comfort zones. However, we do not question the possibility the culture might have kept some diverse talent outside the equation; or what if the culture has kept some of its own employees at bottom due to biases within the organization. The performance management evaluations are themselves be biased and subjective if not structured properly with objectively measurable criteria. Long tenure in an organization is usually considered another metric of success. However, what if the employee has been with the company for more than 10 years because he/she did not want to learn new things and was content with doing the same thing over and over again, or did not get any outside offers all that time because there was not anything particularly successful to grab attention. Usually a successful long tenure in a company means the employee has been promoted during the time or has taken on more responsibility, which is absolutely a sign of success. So, a basic calculation looking at time in an organization without looking at the more nuanced changes should not be the criteria for success. In the same token, gaps in employment should also not be held against an applicant. The applicant might have a disability or another circumstance which required him/her to take time off from work. The machine learning algorithm Amazon had built for its own hiring purposes using its own job applicant data since 2014 had to be scraped by the company when it realized the algorithm was biased. The models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. However, the database was a reflection of the heavy male dominance across the tech industry. In effect, Amazon’s system taught itself male candidates were preferable. The algorithm penalized resumes which included the word “women’s,” as in “women’s chess club captain.” To the company’s credit, it did not keep pushing the use of product when it noticed the bias despite all the investment made on the system. However, it is a good reminder and case study for us when looking at bias in dataset. We need to remember that even when sensitive/protected characteristics (like race, gender, age, etc.) are explicitly ignored in the model, there can still be some data points which can be proxies for these characteristics (zip code, college name, etc..), which can still reflect the same systematic injustices and bias in the dataset. Long story short, predictions based on historical data of a company for a customized tool can further deepen the underrepresentation of females, non-binary applicants,
ethnic minorities, people with disabilities and so on – exactly the type of issue the company wishes to avoid or correct in the first place.

**ASSESSMENT:**
Assessment stage is where applicants are asked to go through different exercises to understand their fit for a certain role. In a traditional sense, the assessment step might include interviews, simulations, case studies, tests or games. The main types of algorithmic assessment tools are focused on facial, speech or emotion analysis during candidates’ interviews or gamified tests on the other. In their research of evaluating the claims and practices of 18 vendors of algorithmic pre-employment assessment, Raghavan (and et al, 2020) cite lack of publicly available information, and lack of information about the validity of these assessments as biggest obstacles to empirically characterizing industry practices. This holds true for most of the assessment algorithms used in the market today.

Inferred traits may not actually have any causal relationship with performance, and at worst, could be entirely circumstantial (Bogen and Rieke, 2018). In other words, the correlations which the algorithms found to build a model, or the traits which the developers built into coding may have nothing to do with a person’s success on the job. So not only we are faced with a black box when it comes to these algorithms (i.e. the workings of the algorithm is not understood or can be explained), but even if we had access to the code and the algorithm itself was explainable, the explanation might not necessarily mean anything.

As Reema Patel, head of public engagement at the Ada Lovelace Institute, puts it “There’s no data that demonstrates that facial recognition technology to profile people works, and effectively, what we’re looking at is a form of pseudoscience that has a potential risk of discriminating against disabled people” (Lee, 2019). The assessment may not work well for people with differences in facial features and expressions if they were not considered when gathering training data and evaluating models; body recognition systems may not work well for a person with disability characterized by body shape, posture, or mobility differences; or analysis tools which attempt to infer emotional state from prosodic features are likely to fail for speakers with atypical prosody, such as people with autism (Guo, Kamar, Vaughan, Wallach, Morris. 2019)

Put aside the fact 1 billion people, or 15% of the world’s population, experience some form of disability according to World Bank and the fact there is not enough work done to solve all the
different biases this population faces, the algorithmic bias in assessment tools does not stop with only the those with disabilities.

EPIC filed a complaint with the FTC alleging that recruiting company HireVue has committed unfair and deceptive practices in violation of the FTC Act. use of micro-expression matching (analyzing the candidate’s facial expressions, their gestures, whether they’re making eye contact, their body language, their speaking speed and the candidate’s choice of words). Yes, HireVue is the most commonly cited example in this category, but it is far from being the only one. Micro-expression matching or analysis also works against those applicants whose native language is different than the language used in the tool; or the facial analysis systems struggle to read the faces of women with darker skin (Buolamwini and Gebru. 2018). As a result, the system either filters out all these candidates either as not fit for hiring, or erroneously flags their data as invalid outliers.

Vendors like Faception, a facial personality analytics tool, suggests their proprietary computer vision and machine learning technology can profile people and reveal their personality based only on their facial image; claiming they can tell if a person has a high IQ, or is more likely an academic researcher, or terrorist. I will constrain from myself from going in a deep dive argument of what sounds like phrenology, a Lombroso-ist approach and the whole unscientific and malevolent aspects of this approach. However, it does raise a red flag because this vendor also lists smart cities, recruitment, retail and insurance in its product verticals.

**SOCIAL PROFILE AGGREGATION:**

Let’s say a candidate has gone through all these stages and is shortlisted for a job offer. Despite the fact a number of states ban employers from looking at candidate’s social profiles to get more information, not all states or countries do. A number of algorithmic tools can now scrape all your social profiles and post on the internet and make recommendations about you to employers by classifying you in certain categories. Michal Kosinski and colleagues have shown machine learning algorithms can predict scores on well-established psychometric tests using Facebook “likes” as data input which are the digital equivalent of identity claims: “Likes” tell others about our values, attitudes, interests, and preferences (Kosinski, Stillwell & Graepel, 2013). On a separate note, as Duarte et al suggest these tools using natural language processing technology “have limited ability to parse the nuanced meaning of human communication, or to detect the intent or motivation of the speaker.” Definitions of what constitutes toxic or concerning content are often vague and highly subjective. (Duarte, Llanso, and Loup, 2017)
In a world where our digital footprint becomes our twin persona and where almost everyone can get their hands on our information, the democratic process and our ability to openly share your views on different issues also comes under pressure. You might not want to take a stand on important societal issues if you know a future employer may make an adverse decision on your employment because of what they saw. Background checks can also surface details about an applicant’s race, sexual identity, disability, pregnancy, or health status, which employers should not consider during the hiring process. Employers should not sacrifice the integrity of the recruitment process in an effort to catch a handful extreme cases of unacceptable behavior. The benefit does not justify the impact on free speech.

CONCLUSION:
There are certainly great opportunities to use AI to analyze a company’s structure and see potential issues with imbalances across employee population, underrepresentation of different groups across various processes, etc; or use AI in a responsible manner to improve your processes. Algorithmic bias may exist even when there is no discriminatory intent on part of the vendor if there if the data was not good, and no employer invests in a product solely to cut costs if they know there might be certain bias and even discrimination issues. However, blindly onboarding with a software without doing a deep dive due diligence is also not a responsible way of conducting business either.

Algorithms are not independent of their developers, nor is the data of the populations upon which they are built without the potential of embedded bias. It is not enough for companies to self-govern their products when the stakes are high. We need better governance mechanisms to be able to hold vendors accountable in more effective ways. We need policies and regulations in place to fight structural injustices and transform societies through fair access to opportunities.
REFERENCES:
Dastin, Jeffrey, Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. 2018 https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scrap-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Merve Hickok @ Aethicist.org


Schulte, Julius. *AI-assisted recruitment is biased. Here’s how to make it more fair*. World Economic Forum. 2019 [https://www.weforum.org/agenda/2019/05/ai-assisted-recruitment-is-biased-heres-how-to-beat-it/](https://www.weforum.org/agenda/2019/05/ai-assisted-recruitment-is-biased-heres-how-to-beat-it/)


Vincent, James, *Google's algorithms advertise higher paying jobs to more men than women*. The Verge. 2015 [https://www.theverge.com/2015/7/7/8905037/google-ad-discrimination-adfisher](https://www.theverge.com/2015/7/7/8905037/google-ad-discrimination-adfisher)