



Introducing

# FAIR<sup>TM</sup>

A global standard for  
the responsible use of  
AI in recruitment.

Dr Buddhi Jayatilleke  
Barbara Hyman

# Contents

Introduction	3
Understanding Bias in Ai	4
The FAIR™ Framework	6
Unbiased	7
Explainable	7
Valid	8
Inclusive	8
Extending FAIR™ to achieve trust	9
Sapia's adherence to FAIR™	10
References	13
About the Authors	14

# Introduction

Ai (Artificial Intelligence) technology is poised to transform every industry, just as electricity did 100 years ago. Between now and 2030, it will create an estimated \$13 trillion of GDP growth<sup>1</sup>. In recent years, HR and recruitment technology has become dense with Ai products. Most CHRO's inboxes are overwhelmed with emails about new solutions. There's big hairy audacious claims of ROI and liberal use of the latest buzzwords for what are often simple and unsophisticated matching tools.

At the same time, there is a growing awareness of the risk in using some Ai technology amidst news articles around algorithmic and automation bias. There is room for valid scepticism with an absence of any form of accreditation of vendors, who often use new scientific approaches and claims that are unpublished, and lack scientific scrutiny. Regulation is light years behind tech innovation. As the market gets denser with new products, so does the rhetoric around the dangers of Ai. In the HR industry a lot of these fears centre around the amplification and automation of human biases via Ai. This is valid, but it also ignores the power of Ai (aka data) to identify and mitigate bias if used wisely. Fear is limiting our capacity for real change.

If you are committed to a culture of decision-making with data and not decision-making from "gut instinct", then Ai literacy and empowerment need to be prioritised in your organisation. This resistance to Ai has happened at the same time the spotlight is on bias interruption in our organisations and institutions. The campaign for racial justice and equality has been amplified by the Black Lives Matter movement.

The right Ai tool can remove bias from your recruitment process and deliver a more diverse workforce. The right data disperses the burden of ignorance inside a company, and can transform your culture. It can do this more effectively than rounds of unconscious bias training which research has shown does not work to change attitudes. This finding has led the UK government to defund all such training.<sup>2</sup> There is no shortcut to making the process of Ai literacy easy for CHROs. The bar must be held high when you are making life changing decisions on the basis of data.

## The Answer

1. **Self-education:** Something this paper is designed to help you with.
2. **Self-regulation:** Thorough impact assessments looking at the holistic candidate experience not just the algorithmic components overseen by joint team comprising HR, legal and cyber security.
3. **Support:** This should be in the form of a guiding framework for making the right decisions

This paper offers a guiding framework Fair Ai for Recruitment (FAIR) centred on a close examination of what constitutes 'fair' and the additional steps to ensuring trust in the technology system which takes into account aspects of the technology vendor organisation and its own systems for transparency and bias mitigation. Ai can deliver powerful and better outcomes for recruiters and candidates, but we must ensure that all recruiting Ai is fair. In the following pages you'll learn how to do that.

---

1 <https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-Ai-frontier-modeling-the-impact-of-Ai-on-the-world-economy>

2 <https://www.bbc.co.uk/news/education-55309923>

# Understanding Bias in AI

## How does bias arise?

In an AI system embedded in a live environment, bias can originate at three key points: data, algorithms and user interaction (See figure 1). While this is a simplified view, it helps structure our understanding and investigation of bias.

A typical AI system uses data and algorithms to model a real-world environment to come up with predictive outcomes that help solve a problem. For example, in recruitment, an AI system could model a candidate based on data in their resume to rank them using some success criteria learnt from past performers.

## Data

Data informs the machine learning algorithm and is the only way it can learn about the environment. The assumptions made by the designer and what data are selected to represent the environment can significantly influence what biases are allowed into the algorithm. Some examples of data used in recruitment include resumés, video, demographic data, personality test outcomes and performance data such as manager ratings. Data is the most common reason for biased AI, a phenomenon known as “garbage in, garbage out”.

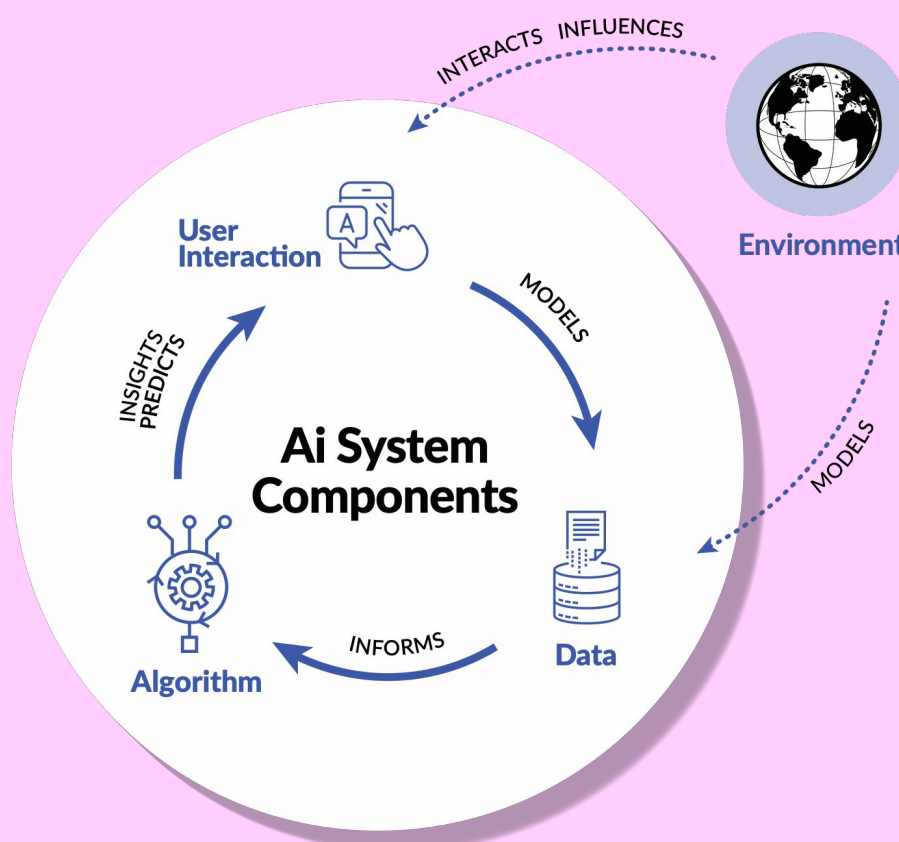


Figure 1: Ai System Information Flow

If data generated using a process plagued with unconscious biases are used in training models, the resulting models can learn and reflect these flaws in a way that leads to inequitable outcomes. It is obvious how data such as resumés, and video can amplify bias.

A good example of this is the Amazon experiment where they built predictive models to screen candidates for technical roles using resumés as input<sup>3</sup>. Their models were trained on resumés submitted to the company and hired over a 10-year period, mostly from men, a reflection of male dominance across the tech industry. As a result, the Ai taught itself characteristics of male resumés as preferred and penalized resumés that included terms indicative of female candidates such as “women’s,” as in “women’s chess club captain”. It even downgraded graduates of two all-women’s colleges without explicit mention of the names of the schools in the resumés.

## Algorithm

Algorithm is the mechanism by which the patterns in the data are discovered and turned into predictions. Algorithms vary in complexity of how inferences are made, ranging from simple linear models to decision-tree based ones, to more complex deep neural networks.

Each algorithm comes with assumptions of what patterns can exist in the data and attempts to maximise some success criteria. While complex algorithms such as deep learning models may provide higher accuracy, their outcomes are harder to explain. If the selection of the algorithm is based mainly on measures of accuracy, and not on fairness and explainability, the resulting models can lead to biased outcomes.

## User interaction

User interaction is how a user interacts with an Ai based system, consuming its outcomes and generating data for future model building. The design of the user interaction can impact the inclusivity of users and the impact from the Ai outcomes.

For example, an interface that is not accessible to people who are visually impaired or who have Dyslexia may filter out data from those sub-groups, making the resulting dataset less representative of the population you are interested in modelling.

Another example is automation bias, where users disregard contradictory information or don’t challenge the outcomes of a computer-generated solution, especially in time-critical domains. For example, recruiters might rely solely on the ranking of candidates without considering how nuanced the ranking is or exploring the reasons behind the ranking.

An Ai system built with the knowledge of these biases needs to have in place measures to test and mitigate them or state clearly the assumptions made in the process to enable fair application of the model. While this is mainly useful for the developers of Ai, users and buyers of Ai must be aware of these biases in order to assess or ask the right questions when using or procuring Ai based tools. Investigation of bias in machine learning models is a relatively new and active area of research (given applied machine learning itself is a nascent technology) and you can find an in-depth analysis in [1].

In the next section we explore the FAIR™ framework, that provides guidelines on how to assess the above biases and build fairer Ai systems.

---

<sup>3</sup> <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scrap-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>

# The FAIR™ Framework

Fairness is the absence of any prejudice or favouritism toward an individual or a group based on their protected attributes in the interactive and non-interactive components of the predictive system engaged in assessing a candidate using a suitable and acceptable metric.

The core focus of the FAIR™ framework is to establish a data-driven approach to fairness that provides an objective pathway for evaluating, challenging and enhancing fairness considerations.

The Fair AI for Recruitment (FAIR™) framework presents a set of measures and guidelines to implement and maintain fairness in AI based candidate selection tools. It does not dictate how AI algorithms must be built as these are constantly evolving. Instead, it seeks to provide a set of measures and guidelines that both AI developers and users can adopt to ensure the resulting system has factored in fairness.

For AI developers FAIR™ provides a set of measures that can be used to demonstrate the fairness of the resulting AI.

For hiring managers and organisations, it provides an assurance as well as a template to query fairness related metrics of AI recruitment tools.

For candidates, FAIR™ ensures that they are using a system built with fairness as a key performance metric.

The core focus of the FAIR™ framework is to establish a data-driven approach to fairness that provides an objective pathway for evaluating, challenging and enhancing fairness considerations. We have based the current version on the concepts around fairness, bias and bias mitigation in AI systems found in research literature listed under References.

Fairness is a complex topic with many contextual nuances related to social and individual circumstances, and in turn, we expect FAIR™ to evolve. We see it as a fluid framework that will evolve as the body of work under ethical use of AI grows.

## What is fair in FAIR™?

It is important to state what FAIR™ considers to be “fair” in the context of recruitment. This is what the quantifiable aspects of the framework are built on. Following is a working definition we have adopted. We have extended the definition found in [1], by not limiting fairness to the predictive outcomes of AI. We see fairness as an end-to-end system consideration related to UI design, data schema design, feature engineering (i.e., what goes in as input), machine learning model training, selection, monitoring and user documentation.

Secondly, we expect the “absence of any prejudice or favouritism” to be defined as a measurable entity using a suitable and acceptable metric.



## Measuring Fairness

In order to demonstrate that an AI system adheres to the fairness definition, FAIR™ expects it to demonstrate four properties, namely: unbiased, valid, explainable and inclusive (see figure 2). These four properties are selected to cover the key themes around fairness found in relevant research (see References).

FAIR™ does not prescribe what measures are to be selected for each of the properties and leaves that choice to the developers. FAIR™ only requires the selected measures to be reasonable in demonstrating each of the properties and acceptable under best practices, legal and specific organisational requirements.

For example, in the case of bias, the 4/5th rule is a possible choice given it is a bias test recommended by the EEOC in the US. In the section under “Sapia’s Adherence to FAIR™” we list the measures adopted by us under each property and provide a good starting point for anyone interested in suitable measures.



Figure 2: FAIR™ Properties

## Unbiased

Outcome from an applicable set of bias tests to demonstrate that the AI is not showing a bias towards a group defined by a protected attribute. We recommend testing training data and predictive model outcomes at both training and live inference stages. At minimum the model’s outcomes must be tested for bias on demographic groups of interest (e.g., gender, race, age group etc).

We do not prescribe a specific bias test, but refer to applicable bias tests listed in literature, such as [1]. See the following section for tests used by Sapia. Interested readers can find comprehensive guides to bias testing including toolkits under IBM Fairness 360 [2] and Aequitas framework from the Center for Data Science and Public Policy at the University of Chicago [3].

## Explainable

Supported by documentation and tools that help interpret the outcomes of the AI solution. We consider explainability at three levels:

**1. Science behind the predictive models:** The research, theories, assumptions, data etc related to building the predictive model. Where possible vendors should publish their approaches for peer review. We do see the challenge in exposing proprietary information related to intellectual property behind predictive models, but on an holistic view, the benefits outweigh the risks. Vendors should be more open about the science behind their products beyond simple descriptions keeping the models a total “black box”.

**2. Interpretation of individual outcomes:** Providing insights to both candidates and hiring managers beyond a single predictive score or label, helping them understand what the AI has learnt about the candidate in making a prediction. These explanations can be simplified to a level that is helpful to the candidates and hiring managers. However, vendors must have the capacity to explain each prediction at a more technical deeper level using methods such as LIME [4].

**3. Transparency with regard to model performance:** Model performance metrics such as precision, recall, mean squared error (MSE), error rates on demographic groups etc should be made available to interested parties. Live model performance data on whether the model is behaving as expected should be monitored and made available to decision makers.

## Valid

As predictive applications, outcomes of Ai need to demonstrate validity, specifically criterion validity. In other words, evidence must be provided on how well the Ai is able to predict a pre-defined measure.

Typically, the machine learning model training process includes a built-in step to establish concurrent validity (an aspect of criterion validity) as it tests each resulting model on an independent data set, not used in training, to validate the model's performance.

Model performance metrics such as accuracy, precision, recall, F1 score in classification models and r2 in regression models are examples of these. The other aspect of criterion validity is, predictive validity, which refers to how accurately the Ai outcome predicts what it is supposed to predict.

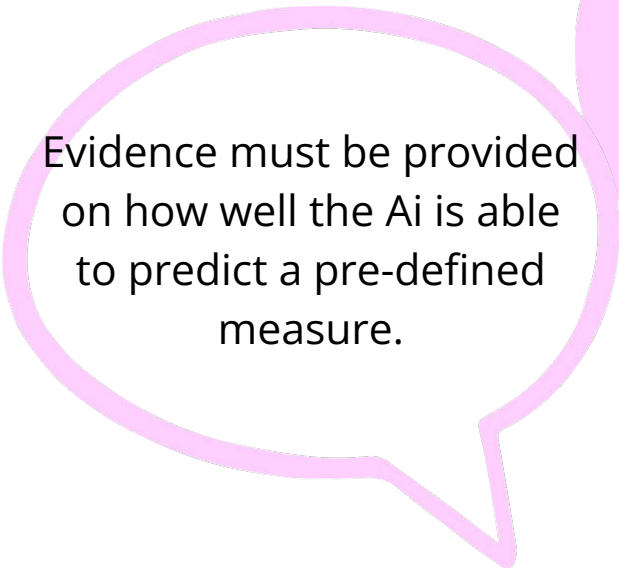
Other forms of validity such as face, construct and content validity can also be demonstrated.

## Inclusive

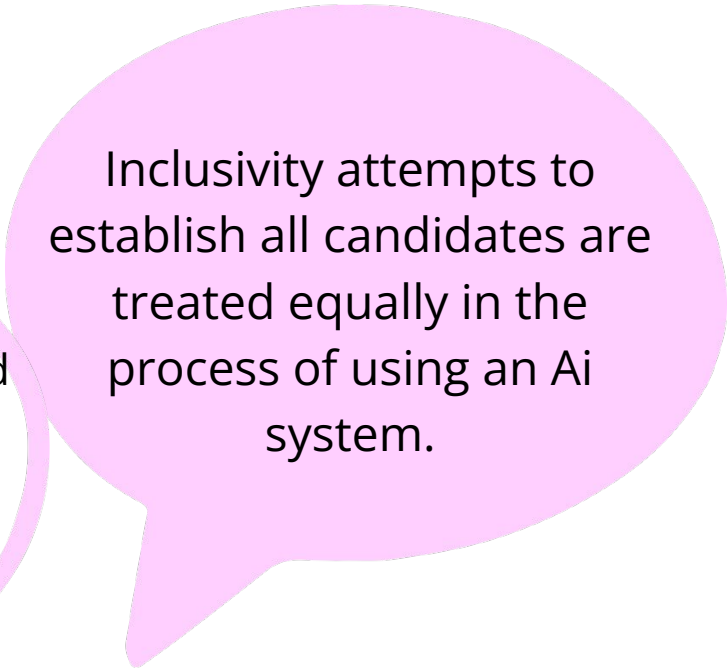
The measure of inclusivity attempts to establish that all candidates are treated equally in the process of using an Ai system. This is the most difficult measure to establish given the subjective nature of "equal treatment".

For example, are timed tests fair for candidates with cognitive disabilities? The important point is that by having inclusivity as a key measure, vendors must consider this as part of the system design, beyond the Ai model.

What measures can be used to demonstrate inclusivity is open to vendors to define and justify. One option is to use user experience metrics such as satisfaction scores, time to complete (an assessment), candidate dropout rates, candidate bounce rates and feedback comments. It can for example show that dropout rates are similar between male and female candidates and that average satisfaction scores are not significantly different between female and male candidates.



Evidence must be provided on how well the Ai is able to predict a pre-defined measure.



Inclusivity attempts to establish all candidates are treated equally in the process of using an Ai system.



# Extending FAIR™ to achieve trust

While fairness is a necessary condition in building trust, it alone is not sufficient to achieve it.

Building a relationship of trust between the operators of Ai and its users is essential in making the users feel safe in using an Ai based tool. While fairness is a necessary condition in building trust, it alone is not sufficient to achieve it. Organisations can extend FAIR™ to achieve trust by demonstrating three more properties (see figure 3).

## Data Privacy and Security

This refers to the measures put in place to protect the data and user privacy, as an organisation. Adherence to regulations such as GDPR and complying to standards such as ISO 27001 and SOC 2 are common ways to establish this.

Moreover, not capturing protected attributes such as gender, race, age etc and not scraping public domain data about users are also good ways to avoid privacy concerns and build trust with candidates.

Literacy in the organisation as a whole around privacy and data protection further strengthens this.

## Team Diversity

It is important to have a diverse team behind the development and management of the product and services surrounding the Ai system. Is the team behind the Ai system reflective of the population of candidates using the system?

A diverse team enables different points of view to be considered internally, especially around the inclusivity of the system, before it reaches end users.

## Transparency

The degree to which the organisation is open about its technology. The explainability listed under FAIR™ is related to transparency from the point of view of the Ai.

Broader transparency around helping candidates understand what is expected of them, company policies and diversity etc is important in building trust.

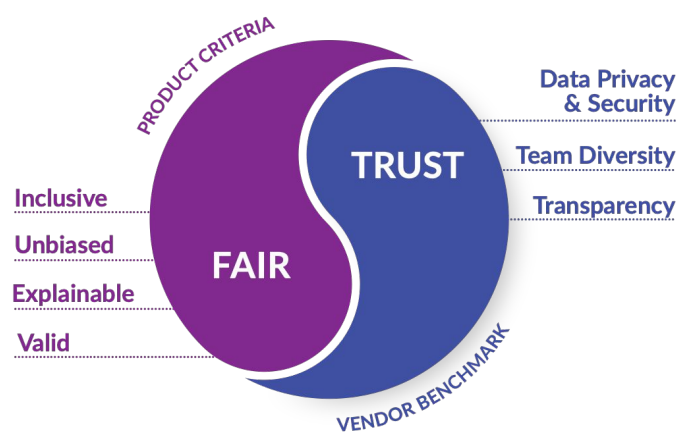


Figure 3: Achieving Trust

# Sapia's adherence to FAIR™

In this section we provide an example by briefly discussing how Sapia adheres to the FAIR™ framework.

## Unbiased

To start with we do not use any demographic attributes, or any other data taken without the consent of the candidate in our models (in that we achieve “fairness through unawareness” as discussed in [1]).

Secondly, we conduct comprehensive testing to uncover any bias in both data we use to train our Ai and the resulting predictive models. The testing happens in the three key stages of the model development and use (See figure 4).

Our default testing looks for adverse impacts on gender and race groups. We do not collect these protected attributes directly from candidates, but use an external service called NamSor, (<https://www.namsor.com/>), to derive race and gender from candidate names.

NamSor is one of the leaders in name to gender, ethnicity and origin classifications. Features in the training data and model outcome scores are tested using effect size, t-test, ANOVA and 4/5th rule (where applicable) across gender and race groups.

We also conduct error rate parity tests across groups to establish that a classifier is making similar errors between groups, for example false omission rate and false discovery rate between male and female candidates. Models that do not pass the test criteria are not deployed.]

The live models are also monitored for adverse impact across gender and race groups. Our “Discover Insights (Di)” dashboard provides live diversity data throughout the applicant funnel for employers. Figure 5 shows a sample of one of the graphs in Di showing the selection rate for gender across applied, recommended (by Ai) and Hired (human decision) funnel.

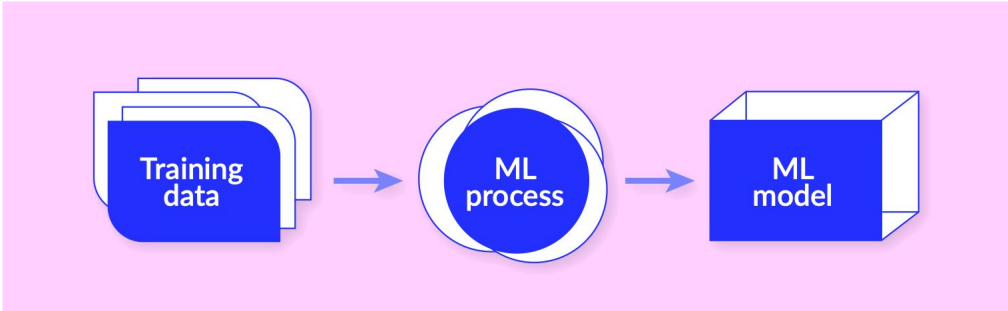


Figure 4: Simplified Model Building Process

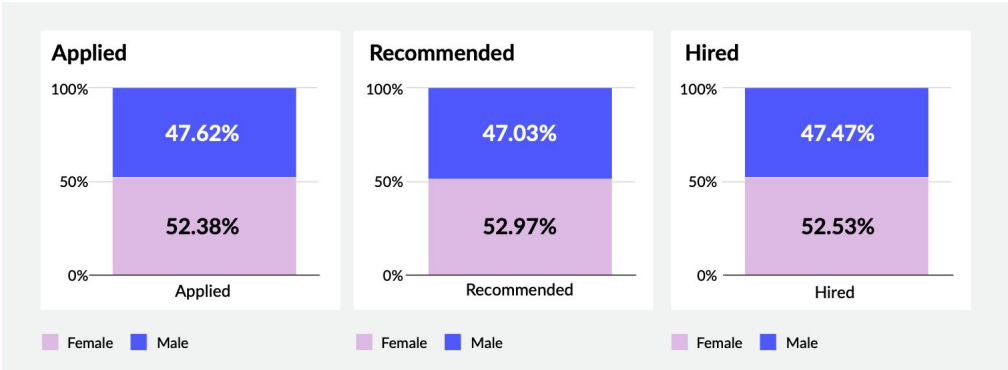


Figure 5: Discover Insights (Di) dashboard

We conduct further independent studies with interested hiring organisations to test whether our algorithms impose any adverse impact on protected attribute groups beyond gender and race.

For example, with one customer we conducted a detailed adverse impact study on gender, race, age groups and English as Second Language (EASL), across the applicant selection funnel.

We also conducted a large study (N > 1,000) on the Amazon Mechanical Turk Platform to test for adverse impact on age, education level, EASL, disability statuses, medical conditions such as Autism Spectrum Disorder and Dyslexia. Please reach out if you are interested in seeing the full adverse impact study report and more details of our standard bias testing.

## Valid

As part of the machine learning process concurrent validity is measured on the test data set aside and not used in training the model. Predictive validity of the models are measured using hiring data. In other words, what percentage of the hiring outcomes come from recommended vs other candidates and the average predictive score difference between hired vs rejected candidates.

## Explainable

We implement explainability on three different fronts.

For the candidate we provide what we call “MyInsights (Mi)”, a personality analysis based on what the Ai has learnt about the candidate. The candidate is then able to provide feedback on whether they agree with the insights or not. An overwhelming percentage (>85% of feedback senders) agree with the Mi report. We also provide candidates with online documentation that describes how the Ai system scores their answers.

For the hiring organisation we provide “TalentInsights (Ti)”, a quantitative insights report listing the underlying input values and benchmarks to help recruiters demystify the final predictive score for each candidate. See figure 6 below for some of the insights available in the Ti report.

Table 2.1 Metrics Sapia uses to establish validity

1	<b>ACCURACY</b> Ratio of correct predictions	3	<b>RECALL</b> Ratio of true positive predictions over all positive candidates in the dataset (true positive + false negative)	5	<b>AREA UNDER THE CURVE (AUC)</b> Area under the curve in the Receiver operating characteristic (ROC) graph. AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one, which is the desired behaviour of a classification model.
2	<b>PRECISION</b> Ratio of true positive predictions over all positive predictions (true positive + false positive)	4	<b>F1 MEASURE</b> A harmonic mean of precision and recall		

For the model developers, LIME [4] based predictions explainer service is available to interpret outcomes at the raw feature level. We refrain from using deep neural network algorithms for building the final predictive models given the challenges in explaining their outcomes.

We also publish our research work, where applicable, to seek peer researcher reviews and share the science behind our Ai. For an example, see our publication on “Predicting Personality Using Answers to Open-Ended Interview Questions” on the journal IEEE Access<sup>4</sup>.

**Inclusive**

In order to measure whether the application experience is similar across different demographic groups, we measure and conduct statistical tests to assess significant differences. The measures we use include:

1. Candidate satisfaction rating: A score between 1-10 given at the completion of the assessment by the candidate.
  2. Candidate satisfaction comment: Along with the score, a candidate can optionally leave a comment. We calculate an engagement rate based on the percentage of candidates who leave a comment. We also infer the sentiment of the feedback which we use as a satisfaction measure.
  3. Time to complete: How long a candidate took to complete the assessment.
  4. Dropout rate: Percentage of candidates who started but did not complete the assessment.
- We look at the difference between gender and race groups for the above measures to discover potential user experiences issues. Each hiring organisations’ Discover Insights (Di) dashboard shows real-time values for the above metrics and comments.

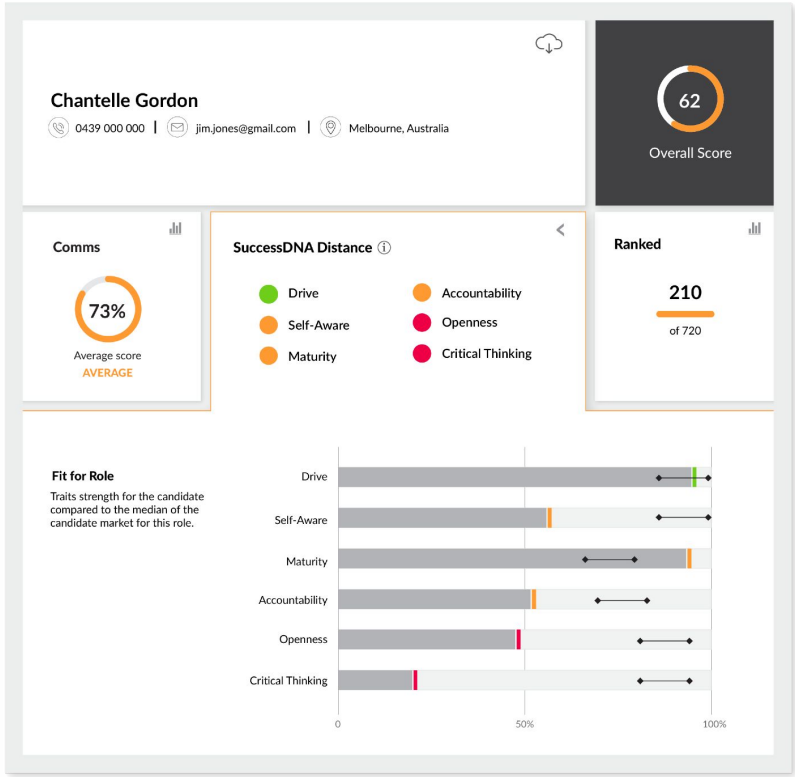


Figure 6: Talent Insights (Ti) dashboard

4 <https://ieeexplore.ieee.org/document/9121971>

# References

[1] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. ArXiv, abs/1908.09635.

[2] IBM's Ai Fairness 360 toolkit (AiF360)  
(<https://developer.ibm.com/technologies/artificial-intelligence/projects/ai-fairness-360/>)

[3] Aequitas framework from the Center for Data Science and Public Policy at the University of Chicago  
(<http://www.datasciencepublicpolicy.org/projects/aequitas/>)

[4] Local Interpretable Model-Agnostic Explanations (lime), <https://lime-ml.readthedocs.io/en/latest/>

## Other Useful References

[5] Sánchez-Monedero, Javier and Dencik, Lina and Edwards, Lilian, What Does It Mean to 'Solve' the Problem of Discrimination in Hiring? (October 2, 2019). Available at SSRN: <https://ssrn.com/abstract=3463141> or <http://dx.doi.org/10.2139/ssrn.3463141>.

[6] Miranda Bogen and Rieke Aaron. 2018. Help Wanted: An Exploration of Hiring Algorithms, Equity and Bias. Technical Report. Upturn.

[7] Tambe P, Cappelli P, Yakubovich V. Artificial Intelligence in Human Resources Management: Challenges and a Path Forward. California Management Review. 2019;61(4):15-42. doi:10.1177/0008125619867910

[8] 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 (FAT\* '20). Association for Computing Machinery, New York, NY, USA, 469–481. DOI: <https://doi.org/10.1145/3351095.3372828>

[9] Michael Kearns and Aaron Roth. 2019. The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press, Inc., USA.

[10] Centre for Data Ethics and Innovation (Part of Department for Digital, Culture, Media & Sport of the Government of UK), Review into bias in algorithmic decision-making (Nov2020), [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/949383/CD-EI\\_review\\_into\\_bias\\_in\\_algorithmic\\_decision-making.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/949383/CD-EI_review_into_bias_in_algorithmic_decision-making.pdf)

# About the Authors



## **Dr Buddhi Jayatilleke**

*Principal Data Scientist, Sapia*

Buddhi is a Data Science leader with a diverse background in machine learning, software engineering and academic research. Buddhi's academic involvements include teaching and senior research fellow positions in leading Australian Universities working on interdisciplinary research projects exploring data driven decision making. Prior to joining Sapia he was the Lead Data Scientist at Culture Amp, where he led various data science projects focused on people analytics. At Sapia, he brings expertise in engineering and data science to help build tools that support fairer candidate selection. Buddhi holds a PhD in Computer Science from the RMIT University and a Master of Software Systems Engineering from the University of Melbourne.



## **Barbara Hyman**

*CEO, Sapia*

Barb started her career as a solicitor. She quickly realised her skills and personality were better suited to business and after completing an MBA, she spent a decade at BCG in consulting and later as Head of HR & Marketing for the Australia & NZ region. It was that role followed by her CHRO experience at the largest digital company in Australia – the REA Group, that inspired her to find a better way to make hiring and promotion fairer, more human and way less taxing for the organisation. Barb holds a BA/ LLB (Hons) from Monash University and an MBA from the University of Melbourne.

In less than 3 years, Sapia has become one of the most trusted mobile-first Ai recruitment platforms, used by ASX and FTSE listed companies, with a candidate every two minutes in any one of 34 countries around the world engaging with their unique Ai chatbot Phai. What makes their approach unique it's disruption of three paradigms in recruitment -candidates being ghosted, biased hiring and the false notion that automation diminishes the human experience.

By asking only 5 behavioural questions relevant to the role, taking the candidate 15- 20 minutes, they can extract up to 80 features that determine suitability for a role. Candidates respond in their own time with every candidate receiving personalised feedback, with coaching tips. No sensitive information is captured like gender, age and race. The Ai that sits behind the interview only uses the textual answers to calculate a "suitability score".

The Ai solution is built using principles of structured interviews, personality theory, natural language processing (NLP) and machine learning. The team believe that transparency drives trust and so have published their research, and adverse impact testing. The end result for companies – bias is interrupted at the top of the funnel, inclusivity is enhanced since chat is intuitive and trusted by most, and your hired profile starts to look more like your applicant profile. No one misses out, and your team never miss out on latent talent. Your hiring managers make more objective decisions, empowered by Phai their co-pilot, with data driven profiles and interview questions to draw from. With a winning candidate experience and go live in less than 24 hours, how can you not use their unique technology to hire with heart (and smarts).



# Cut through the bias, with a greater reach, and transform the way you hire.

Sapia is a frontier interview automation solution that solves three pain points in recruiting – bias, candidate experience, and efficiency.

Customers are typically those that receive an enormous number of applications and are dissatisfied with how much collective time is spent hiring.

Unlike other forms of assessments which feel confrontational or irrelevant Sapia's FirstInterview™ is built on a text-based conversation which is totally familiar because text is central to our everyday lives.

Every candidate gets a chance at an interview by answering five relatable questions.

Every candidate also receives personalised feedback (99% CSAT). Ai reads candidates' answers for best-fit, translating answers into personality readings, work-based traits and communication skills.

Candidates are scored and ranked in real-time, making screening 90% faster.

Sapia fits seamlessly into your HR tech-stack, and with it, you will get 'off the Richter' efficiency, reduce bias and humanise the application process.

Hire with heart.

The logo for Sapia.ai features the word "sapia" in a bold, lowercase, sans-serif font. A small pink smile-like curve is positioned under the letter "i". To the right of "sapia" is ".ai" in a smaller, lighter font.