Empowering HR: How Arena's AI Solutions Mitigate Bias to Build High Performing Workforces

Patrick Hagerty, Chief Data Scientist, Arena Analytics

Introduction

Artificial Intelligence (AI) technology is rapidly evolving with a continuous stream of new applications seeking to improve all aspects of our personal and professional lives. In particular, AI has demonstrated enormous potential in workforce management, including how we recruit, hire, and retain talent.

Employee retention is a major economic issue for organizations. In 2023 alone, voluntary resignations accounted for a staggering 60%[1] of all job separations. The resultant costs of recruiting and training new employees was significant, with estimates[2] as high as 2.5 times an employee's annual salary. To improve talent management, organizations are turning to the unparalleled predictive analytics of AI to supplement human decision-making. AI can uncover correlations that reliably predict how long an applicant is likely to stay in a specific role, empowering organizations to make better hires, create a more diverse workforce, and drastically reduce the costs associated with hiring and training new employees.

However, the transformative potential of AI to streamline workforce management processes is not without risks. AI can potentially reinforce human bias in talent acquisition, resulting in discriminatory hiring and the elimination of highly qualified applicants. Amazon's 2014 initiative to use AI for talent acquisition inadvertently built gender bias into Amazon's AI model, which was trained on data from Amazon's historical database of resumes consisting primarily of men. As a result, the AI model was trained using terms commonly associated with male engineers.[3]

This example illustrates the importance of bias mitigation and transparency in your AI model. It highlights the necessity of examining data sources and ensuring that algorithms are trained on

---

[1]Bureau of Labor Statistics, (2023, January 30) *Job Openings and Labor Turnover Summary* (Press Release). https://www.bls.gov/news.release/jolts.nr0.htm

[2]McFeely, S., & Wigert, B. (2019, March 13). *This Fixable Problem Costs U.S. Businesses $1 Trillion*. Gallup. https://www.gallup.com/workplace/247391/fixable-problem-costs-businesses-trillion.aspx

[3] Reuters *Amazon scraps secret AI recruiting tool that shows bias against women*

diverse and representative datasets. By adhering to rigorous standards organizations can mitigate the risk of bias and ethically and responsibly harness the transformative potential of AI. Arena's Adversarial Fairness approach stands out as a proven solution for mitigating bias—both conscious and unconscious—within hiring processes.

Moreover, as governments around the world move to establish frameworks for ethical AI deployment, Arena is well-positioned to support clients in navigating this dynamic regulatory landscape. Our proactive engagement with regulatory agencies, such as the National Institute of Standards and Technology (NIST), demonstrates a commitment to shaping AI policies and ensuring our model complies with evolving regulations.

In this white paper we will examine Arena's Adversarial Fairness approach to bias mitigation. We will explore the critical factors organizations should consider when choosing an AI tool and how Arena's approach ensures accountability and transparency. And we demonstrate the tangible results clients are achieving with Arena's bias mitigation techniques. This paper outlines Arena's methodologies and also emphasizes its commitment to compliance with existing AI-related regulations and the development of equitable and practical regulations in the future.

Ensuring Ethical and Responsible AI

The risk of bias in the hiring process along with concerns for protecting applicants' data have drawn international scrutiny and regulation. The EU has proposed the EU Artificial Intelligence Act,[4] which utilizes a risk-based approach. In the United States the Federal Trade Commission (FTC) is monitoring companies' claims about the capabilities of their AI tools under consumer privacy laws[5]. And the Equal Employment Opportunity Commission (EEOC) is monitoring AI for its impact on protected classes to ensure that AI used in hiring and other employment decisions complies with federal civil rights laws[6].

In addition, individual states also are focused on ensuring trust in AI systems. To date, nine states, California, Colorado, Connecticut, Virginia, Montana, Indiana, Oregon, Tennessee and Texas have laws regulating AI[7]. New York City passed one of the first most far-reaching laws with its 2023

[4] European Commission. (n.d.). *European Approach to Artificial Intelligence*. Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

[5] How the FTC is influencing AI regulation https://www.dataguidance.com/opinion/usa-how-ftc-influencing-ai-regulation#:~:text=More%20recently%2C%20the%20FTC%20also,using%20the%20improperly%20obtained%20data

[6] (n.d.). *Artificial Intelligence and Algorithmic Fairness Initiative*. US Equal Employment Opportunity Commission. https://www.eeoc.gov/ai

[7] (n.d.). *The State of State AI Laws: 2023*. Epic.org. https://epic.org/the-state-of-state-ai-laws-2023/

requirement[8] that all employers conduct third-party bias audits of any algorithm involved in a hiring decision.

Illinois passed The Artificial Intelligence Video Interview Act[9] (the "AI Act"), which placed notice, consent, confidentiality, and data destruction responsibilities on employers who elected to use artificial intelligence (AI) to evaluate job candidates. And, in 2023 Illinois proposed an amendment[10] to the Illinois Human Rights Act, concerning the use of AI in employment, "an employer that uses predictive data analytics in its employment decisions may not consider the applicant's race or zip code when used as a proxy for race to reject an applicant in the context of recruiting, hiring, promotion, renewal of employment, selection for training or apprenticeship, discharge, discipline, tenure or terms, privileges, or conditions of employment."

Arena Connects with Regulatory Agencies to Shape AI Policies

Arena is well positioned to support clients in this dynamic regulatory environment. For over 15 years, we have worked with private and public sector entities to provide guidance on proposed regulations. We provided input to the National Institute of Standards and Technology (NIST) as they developed methodologies for assessing, improving, and ensuring the trustworthiness of AI systems, and the responsible practices for designing, developing, and deploying AI systems[11].

The recent *NIST Special Publication 1270*, *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*[12] defines how organizations should address the risk of bias in AI. It defines three major categories of bias in AI: systemic, statistical, and human. And identifies three challenges for mitigating the bias: datasets, testing and evaluation, and human factors. NIST also introduces preliminary guidance for addressing these challenges.

---

[8] The NYC Department of Consumer and Worker Protection (DCWP. (n.d.). DCWP - Automated Employment Decision Tools (AEDT). New Laws & Rules Automated Employment Decision Tools (AEDT). https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page#:~:text=Local%20Law%20144%20of%202021,audit%20is%20publicly%20available%2C%20and

[9] (n.d.). *Public Act 101-0260*. Illinois General Assembly. https://www.ilga.gov/legislation/publicacts/fulltext.asp?Name=101-0260

[10](2024, April 19). *HB3773 Engrossed AN ACT concerning business.* State Net. https://custom.statenet.com/public/resources.cgi?id=ID:bill:IL2023000H3773&ciq=ncsl&client_md=8f7e95823cba8d3a00c86cd500fc8976&mode=current_text

[11] https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

[12] Schwartz, R. , Vassilev, A. , Greene, K. , Perine, L. , Burt, A. and Hall, P. (2022), Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD, [online], https://doi.org/10.6028/NIST.SP.1270, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=934464 (Accessed March 5, 2024)

- Systemic bias refers to long-standing biases encoded in society over time resulting from practices or norms that favor or disadvantage certain social groups. These biases are present in the datasets used in AI.

- Statistical bias results from errors when the sample is not representative of the population. In AI these biases are present in the datasets and algorithmic processes used in the development of AI applications and often arise when algorithms are trained on one type of data and cannot extrapolate beyond those data.

- Human biases reflect systematic errors in human thought. These biases are often implicit and relate to how an individual or group perceives information when making decisions or filling in missing or unknown information.

Arena Uses Adversarial Fairness to Mitigate Bias

Arena's Adversarial Fairness approach stands out as a proven solution for mitigating bias in hiring processes. By systematically identifying and eliminating latent correlations related to protected classes, Arena's platform enables organizations to meet and exceed the EEOC 4/5 rule while still maintaining the accuracy of retention predictions.

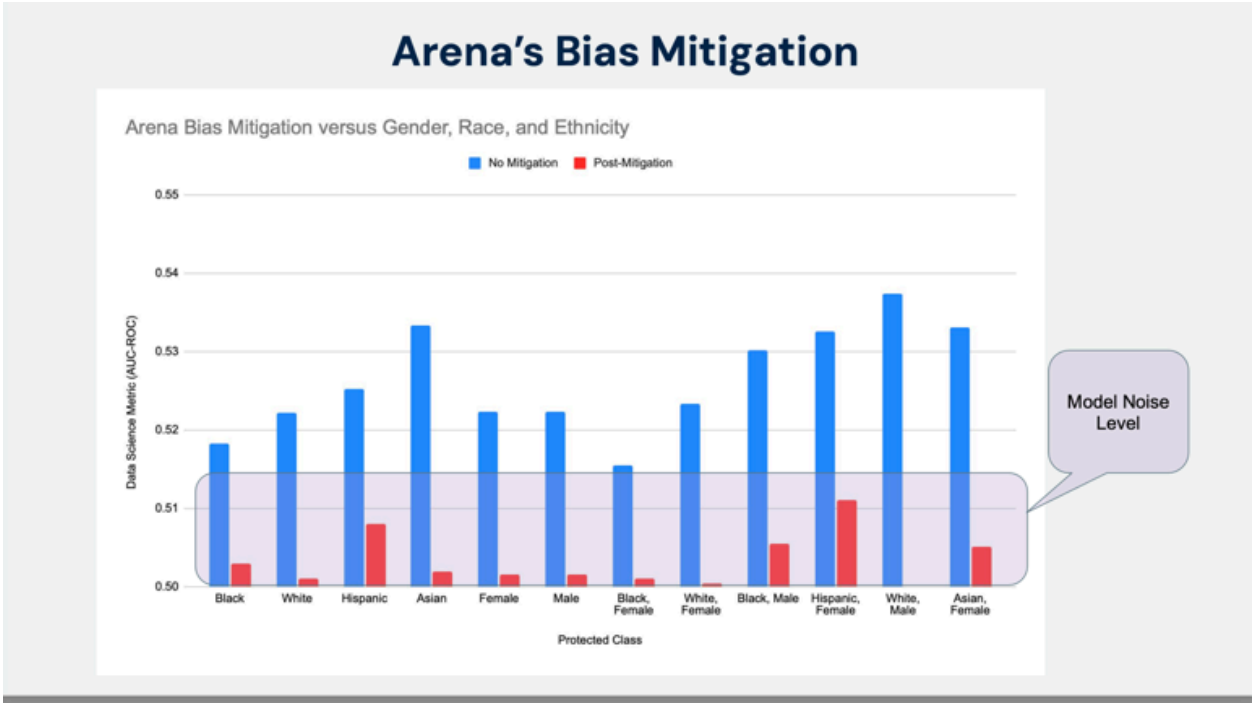


Figure #1. The Impact of Arena's Bias Mitigation

This chart shows the impact of Arena's bias mitigation. On the horizontal axis are the protected classes which we are measuring. The vertical axis identifies how much latent correlation there is about this protected class and the retention prediction.

Without Arena's model, you can see high levels of bias, shown in blue, across all protected classes, raising concerns about compliance with EEOC VII[13] regulations. Arena effectively eliminates bias and still maintains the accuracy of the retention prediction. Bias is reduced to AI noise model levels, which is as good as it can get statistically, plus you maintain the desired predictive accuracy regarding the best person to hire.

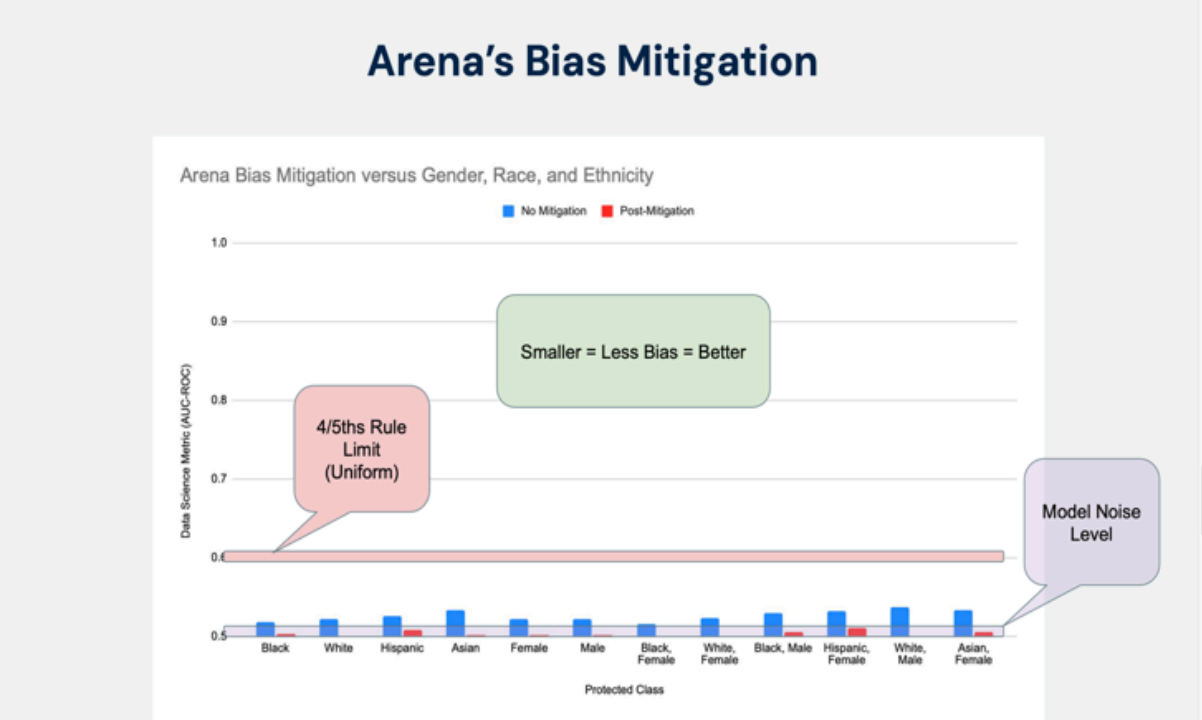


Figure #2. Arena Goes Beyond EEOC Requirements

Using Arena's Adversarial Fairness technique, we remove the obvious correlations and as you can see by our results shown in red, we are well below the 4/5 rule ensuring our clients find the best applicant for each position with the greatest likelihood of being retained. And we are continuously updating our model to extend our bias mitigation capabilities to remove more bias than currently mandated by the EEOC.

[13] (n.d.). *Title VII of the Civil Rights Act of 1964*. US Equal Employment Opportunity Commission. https://www.eeoc.gov/statutes/title-vii-civil-rights-act-1964

It's All About The Data

The significance of data quantity, quality, and diversity cannot be overstated when it comes to mitigating bias during the model design and training processes. Arena's platform exemplifies this approach. We source, analyze, and correlate thousands of data points using input data from government and non-profit organizations and outcome data from our clients. The scope of our data sources enables our platform to identify patterns, remove biases, and make more accurate predictions.

Input Data: Standardization Uncovers Patterns

Arena has compiled more than 10 years of detailed occupational data from the Bureau of Labor Statistics and the U.S. Department of Labor Occupational Information Network (O*NET)[14]. We integrate standardized occupation codes with corresponding occupational details including job descriptions, knowledge base requirements, and skill sets. With this data our AI model can identify patterns that exist in one sector of the labor market and use that knowledge as a starting point for identifying similar patterns in different regions and industries.

Arena's model incorporates extensive data on protected classes, which is paramount for mitigating biases against race, ethnicity, and gender. Our Adversarial Fairness technique represents a major advancement in overcoming bias for marginalized communities.

Outcome Data Is Critical for Overriding Bias in The Hiring Process

Arena also stands apart for the volume of proprietary client outcome data built into our model. At the beginning of 2024, 15% of the US healthcare workforce was on Arena's platform, consisting of 3.1 million healthcare employees and job applicants. And, that number is growing by 600,000 workers per year. This dataset includes compensation data, demographic, and credentialing data.

By harnessing such diverse and robust data, Arena's platform is uniquely equipped to remove the risk of underlying biases while continuously refining the accuracy of our AI models, thereby setting a new standard in AI-driven hiring.

Removing Bias Identifies the Best Applicant And their Potential for Retention

Arena's model predicts applicant retention by analyzing a range of factors and contexts that include the local economy and labor market, competing jobs nearby, the specific position, location, department, and shift that the individual may work. By removing inherent biases our platform can accurately predict which candidates have the greatest likelihood of remaining in their position for an

[14] National Center for O*NET Development. O*NET 28.2 Database. *O*NET Resource Center*. Retrieved March 5, 2024, from https://www.onetcenter.org/database.html

extended period of time. Each applicant receives a Retention Prediction score indicating if they are very likely, likely, or unlikely to be retained over time.

For example, if you are hiring for a dining services coordinator at a senior living facility and one applicant's resume indicates they had only stayed between 1-3 months at a succession of quick service restaurants, human bias might conclude that this applicant would not stay long in a job.

Our platform removes latent correlations about the applicant's previous jobs that are unrelated to the knowledgebase and skills required for the specific Dining Services Coordinator role at the senior living facility. When only considering data that focuses on the requirement of the job, our model predicts that the applicant is likely to be retained in this specific role.

The Importance of Transparency in AI Models

Accountability, according to NIST, is the most critical aspect of governance. And the key to accountability is transparency; defined by NIST as clearly documented AI models along with a designated individual or team responsible for bias mitigation. Most importantly, NIST notes that a central component of risk management for AI bias is the acknowledgement that risk mitigation – not risk avoidance – is the most effective factor in managing risks.

As the Amazon example demonstrates, bias can be introduced into AI models through the data you use to train the model, potentially resulting in discriminatory predictions that impact individuals belonging to protected classes. The EEOC is beginning to address this concern through the recent guidance[15] it issued to employers on how to comply with Title VII regulations as they implement AI platforms.

Arena's Model Card Defines Our Bias Mitigation Model

There are numerous AI tools supporting workforce management. A vendor's model card will help you understand which AI tool aligns with your use case and business goals, ensures data quality and security, and will deliver measurable results. The model card provides transparency as to the purpose of the model, its limitations, expected performance, and potential biases.

Arena's AI model has a three-fold structure, an embedding network, a prediction network, and an adversarial network.

---

[15] *Select Issues: Assessing adverse impact in software, algorithms, and artificial intelligence used in employment selection procedures under Title VII of the Civil Rights Act of 1964*. (2023, May 18). US EEOC. https://www.eeoc.gov/laws/guidance/select-issues-assessing-adverse-impact-software-algorithms-and-artificial

- The embedding network transforms input data into embeddings, which are utilized by the prediction and adversarial networks.
- The prediction network is trained to minimize the prediction error.
- The adversarial network is trained to accurately predict protected classes from the embeddings.

The embedding network is trained with custom orthogonally projected gradients to minimize the adversarial network's prediction error while maximizing the prediction network's accuracy.

Intended Use

Arena's bias mitigation model serves as a preprocessing step to minimize sensitive attributes in the input data, such as race and gender. The model also ensures that we retain the ability to make accurate predictions based on the original input data. Additionally, the model is intended to measure how much information is revealed from the input data about the protected classes.

Metrics

Arena uses the Area Under the Curve (AUC) metric to evaluate the performance of both the prediction network and adversarial network. The AUC offers a comprehensive measure of model performance. We compare the metrics to logistic regression models trained on the same data to validate the effectiveness of our bias mitigation model in reducing bias while maintaining accurate predictions.

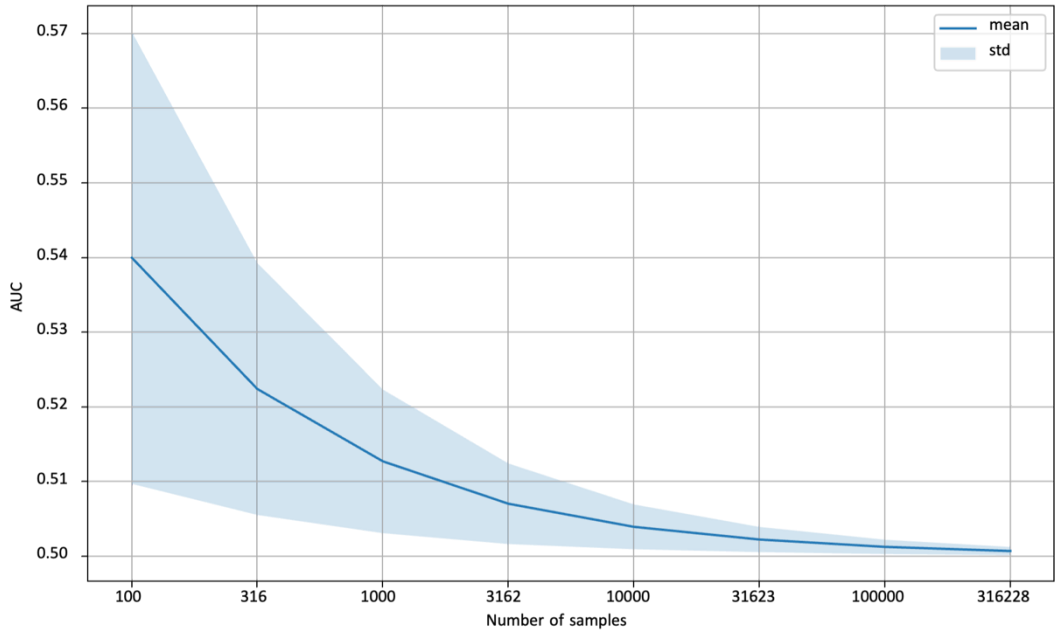


Figure 3: The noise bound of bias mitigation depends on the volume of data. The dark blue plot shows the mean the AUC of random binary classifiers with the shaded region corresponding to +/- one standard deviation. Our training goal to be within the noise threshold depends on the size of the protected class.

During the training process, cross-entropy is used as the loss function for both the prediction network and the adversarial network. Cross-entropy is used in machine learning to assess the difference between predicted probabilities and actual outcome.

Since bias mitigation is a preprocessing step, there is no binary decision made on the output. Instead of accuracy-related metrics that depend on a specified threshold, the AUC metric is a function of all threshold values.

Ethical Considerations

In the Arena model, information relating to an individual's protected class is self-reported without any external verification. The model also may have an imbalance in the volume of data for specific classes which can impact the ability to mitigate bias for those classes. If protected class information is mislabeled, it can extend the time required to train the model and potentially impact the accuracy of the model's predictions. However, we do not expect mislabeling to amplify bias because the output does not contain information associated with the correctly or incorrectly labeled classes.

Training Data

Our model contains data from job applicants who agree to provide their data directly through the Arena platform. The platform has protected class information from 450,000 applicants of Arena clients which is normalized and used to train the embedding model and the adversarial model.

The Arena platform also has labeled performance data on more than three million applicants and one million hires from Arena clients. These inputs also are normalized to train the prediction model.

Caveats and Recommendations

Arena's bias mitigation process only includes protected classes that are specifically enumerated. The model can only identify latent correlations among the input data and the protected classes. Therefore, any classes not in the training data are at risk of bias propagation or amplification. Additionally, bias can be reintroduced with any data features that are added after preprocessing.

Arena's model mitigates bias in a sequential process. The model's performance is dependent on the order of the protected class, especially if there is a lower volume of data for a specific class.

Conclusion

The indispensable role of AI in building diverse workforces cannot be overstated, particularly in today's dynamic and evolving job market. As organizations strive to navigate challenges such as high turnover rates and regulatory scrutiny, Arena's AI platform offers a powerful solution to augment human decision-making and enhance talent management practices.

However, the potential of AI comes with its own set of challenges, most notably the risk of perpetuating bias and discrimination in hiring processes, underscoring the critical importance of transparency, accountability, and ethical responsibility in AI deployment

Arena's AI model incorporates a decade of experience building and refining our AI platform for clients. By embracing ethical AI practices and leveraging innovative solutions like Arena's Adversarial Fairness approach, organizations can unlock the full potential of AI to create inclusive workplaces and drive sustainable growth.



Patrick Hagerty is Chief Data Scientist at Arena Analytics. Prior to joining Arena, Patrick served as the Chief Data Scientist at Adaptive Management, building financial models with alternative data sources. Patrick also served as the Director of Research at In-Q-Tel Labs, leading machine learning efforts in object detection, super/resolution of satellite imagery, attributable synthetic biology, and cybersecurity. Patrick also spent a decade as an Applied Research Mathematician at the National Security Agency developing and optimizing algorithms for quantum computation, cryptanalysis, and high-performance computing.

Glossary

Adversarial Fairness - Arena's platform uses a machine learning technique known as Adversarial Fairness to address the bias in AI algorithms. Adversarial fairness mitigates bias by introducing a discriminator that assesses and penalizes unfair predictions. This ensures the AI model treats all individuals or groups equally, regardless of their underlying characteristics such as age, gender, race, or socioeconomic status.

Artificial Intelligence (AI) - The term, Artificial Intelligence was first introduced at the 1956 *Dartmouth Summer Research Project on Artificial Intelligence.*

Artificial intelligence technology enables computer systems to perform tasks commonly associated with human cognitive functions — such as identifying patterns and making predictions based on existing data. Over the decades, AI has evolved with the disciplines of machine learning, deep learning, and neural networks which extend the ability of AI systems to process massive amounts of data and make increasingly more accurate predictions over time.

Bias Mitigation - Bias mitigation is the process of identifying and eliminating biases that may be present in an artificial intelligence tool. Bias can be introduced through the data that is used to train the models. Bias mitigation helps prevent AI systems from perpetuating and reinforcing discriminatory patterns of behavior.

Ethical AI - AI raises concerns and questions around data privacy, bias, transparency and intellectual property rights. Ethical AI is the development of AI systems that meet legal and ethical standards including insuring data privacy and security and overcoming AI bias and disinformation.

Generative AI - Generative AI systems fall under the category of machine learning. However, while machine learning algorithms analyze data to find patterns and make predictions Generative AI algorithms generate new content including text, images and video in response to a prompt.

NIST Special Publication 1270 - *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. This document by the National Institute of Standards and Technology (NIST) identifies the issues of bias in AI and proposes socio-technical guidelines for addressing and managing these biases.

O*NET - The Occupational Information Network (O*NET) is the nation's primary source of occupational information with data on hundreds of standardized and occupation-specific descriptors. The program was developed under the sponsorship of the U.S. Department of Labor/Employment and Training Administration through a grant to the North Carolina Department of Commerce.