Unmasking AI: My Mission to Protect What Is Human in a World of Machines is an insightful exploration of the ethical and social implications of artificial intelligence by Joy Buolamwini. Through the author's personal experience and research, the book reveals the phenomenon of “coded gaze” in AI technology, i.e., how the potential bias in AI systems maps the structure of discrimination in the real society, such as racial discrimination, sexism, and ability discrimination.

Through multiple examples, the author demonstrates how AI systems fail to do justice to diverse populations due to data and algorithmic limitations. For example, the author describes how a facial recognition system she developed could not recognize dark-skinned faces accurately, which made her aware of racial and gender bias in technology development. The book also discusses the widespread use of AI technology and its harm to society, such as data privacy violations, mass surveillance, employment discrimination, and the exclusionary impact of algorithms on marginalized groups.

At the same time, the author emphasizes the potential of AI and the positive changes it brings. She calls on technology developers to put equity and inclusivity at the center of AI design, while encouraging the public to engage in discussions about the future of AI to ensure that the technology serves all of humanity and does not further exacerbate inequality. Through her story, the author not only opens readers' eyes to the double-edged nature of technology, but also conveys insights into how to balance technological innovation with ethical responsibility.

Working in the field of AI, I have come to the deep realization that technology development is not just a scientific issue, but a social one. We cannot solve these problems with mathematical models or programming techniques alone, and must examine the application and impact of technology from the perspective of humanities and social sciences. For example, whether the selection of training data is comprehensive, whether the algorithms are able to treat different populations fairly, and whether the developers truly understand the needs of the communities they serve.

In my research, I have tried to use an NLP-based model to analyze a set of medical text data to predict a patient's condition grading. However, in real-world applications, I found that the model's performance significantly degraded when processing text containing dialects and slang. After in-depth analysis, I found that this is directly related to the distribution of the training data - the model is mainly trained based on the standard Mandarin corpus, while the dialect corpus accounts for less than 10%. This imbalance in data distribution directly leads to the phenomenon of bias in the practical application of the model, i.e., the model's prediction tends to be inaccurate for patient records expressed in dialects.

This phenomenon impressed me that the “imbalance” in the technology not only affects the applicability of the technology, but also has unfair consequences for certain groups. For example, in healthcare scenarios, this bias could lead to misclassification of patients who speak dialects, which could affect the accuracy of diagnosis and even delay treatment.

Through this experience, I realized that the unfairness of technology is not an isolated incidental phenomenon, but is present throughout the entire process from data collection, algorithm design to system application. As emphasized in the book, data is the cornerstone of AI, and if there is bias in the data, the technology can only perpetuate or even amplify that bias. Therefore, when designing AI systems, we must pay more attention to the diversity and balance of data. Specifically, when collecting data, we should proactively cover samples of various languages, genders, ages and cultural backgrounds. In algorithm design, a bias correction mechanism can be introduced, for example, through a weighted loss function to balance the impact of different groups. In addition, more attention should be paid to the system's performance on vulnerable groups when evaluating the model, rather than just pursuing the improvement of the overall performance index.