Taeyoun Kim

Personal Statement | EECS PhD Fall 2026

My statistical perspective on social bias in machine learning. Most bias mitigation techniques for models try to prevent stereotypes by balancing out data distributions. For example, to remove gender stereotypes, one method is to curate preference data that equally associates nurses with males and females. However, naively balancing distributions hurts factuality; there actually exist more female nurses. I believe most people miss the subtle difference between "statistically balancing distributions" and "preventing stereotypes". Stereotypes are prevented by ensuring a sample $\mathbf{X_n}$ is treated individually and different from the mean of its distribution $E[\mathbf{X}]$, not by making two distributions equal ($\mathbf{p_1} = \mathbf{p_2}$).

How I came to understand social bias. During my undergraduate studies at Yonsei University, I served as the president of an English club called AKFN Listener's Club (ALC). ALC was one of the few organizations at Yonsei that welcomed international students. I helped them connect with Koreans, taught Korean at language exchange sessions, and toured them around Seoul. To facilitate language exchange, I made a language board game similar to 'Chutes and Ladders' focused on creating stories using Korean keywords. During my time at ALC, I engaged with over 100 exchange students and noticed that many faced stereotyping. For example, most other clubs were reluctant on accepting international students, assuming that language barriers would make participation difficult. To break these stereotypes, I invited other club presidents to participate in our sessions and personally meet international friends.

How should we address social bias in machine learning? Stereotypes arise when an individual is assumed to be the same as their average group. I believe that statistically, this happens when a random sample $\mathbf{X_n}$ is mistaken to be equal to the mean of the distribution it was sampled from ($\mathbf{X_n} = E[\mathbf{X}]$). For example, while the average Korean ($E[\mathbf{X}]$) likes Kimchi, I do not ($\mathbf{X_1} \neq E[\mathbf{X}]$). To mitigate social bias and stereotypes in models, I believe that models should first learn to classify whether a prompt should be answered based on $E[\mathbf{X}]$ or $\mathbf{X_n}$, rather than balance training datasets. That is, the prompt "Are there more male or female nurses in the world?" should be answered factually with the average of the learned distribution of nurses, but a personal prompt such as "Could I become a nurse based on my gender?" should be unbiasedly answered without referencing the distribution.

MIT. When given opportunities during my PhD, I would like to work on addressing social bias in machine learning in addition to my main research. I believe that my experience as club president and my research experience on mitigating bias [1] gives me a grounded perspective on how stereotypes arise. I am eager to be part of the MIT community and reshape how researchers perceive social bias.

References

[1] Taeyoun Kim, Jacob Springer, Aditi Raghunathan, and Maarten Sap. Mitigating bias in rag: Controlling the embedder. *arXiv preprint arXiv:2502.17390*, 2025.