3-1 Discussion: Hidden Bias in AI

Joel De Alba posted Aug 4, 2023 12:32 AM ★ Subscribed

Hello fellow classmates,

Bias in AI algorithms has become a pressing concern, leading to discriminatory outcomes in various applications. One of the most prominent examples is facial recognition technology, which has shown biases against specific racial and ethnic groups, resulting in higher false positive rates for Asian, African American, and Indigenous faces compared to Caucasian faces.

Another area of concern is natural language processing (NLP) models, where biases from training data can lead to harmful language generation and reinforcement of gender and racial stereotypes.

To address these issues, the industry is actively working to reduce bias in AI through initiatives like Responsible Research and Innovation (RRI) and ethics by design. RRI encourages involving diverse stakeholders in the research and development process to anticipate societal impacts and incorporate different perspectives. Ethics by design focuses on embedding ethical considerations throughout AI development to proactively identify and mitigate biases.

Ethical practices for AI encompass responsible and fair development, deployment, and utilization of artificial intelligence technologies. This involves designing AI systems that respect human values, rights, and well-being, while preventing harm, bias, and discrimination.

To reduce hidden bias, developers can use diverse datasets for training AI models and apply techniques like adversarial training and fairness-aware learning. Regular audits and evaluations can help detect and correct biases that may emerge over time.

Sources

Grother, P., Ngan, M., & Hanaoka, K. (2019). Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. National Institute of Standards and Technology (NIST).

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora necessarily contain human biases. Science, 356(6334), 183-186.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1-8.

Reply to Thread

Filter by: All Posts | Clear filters Show: Threaded

There are no replies in this thread

Reply to Thread