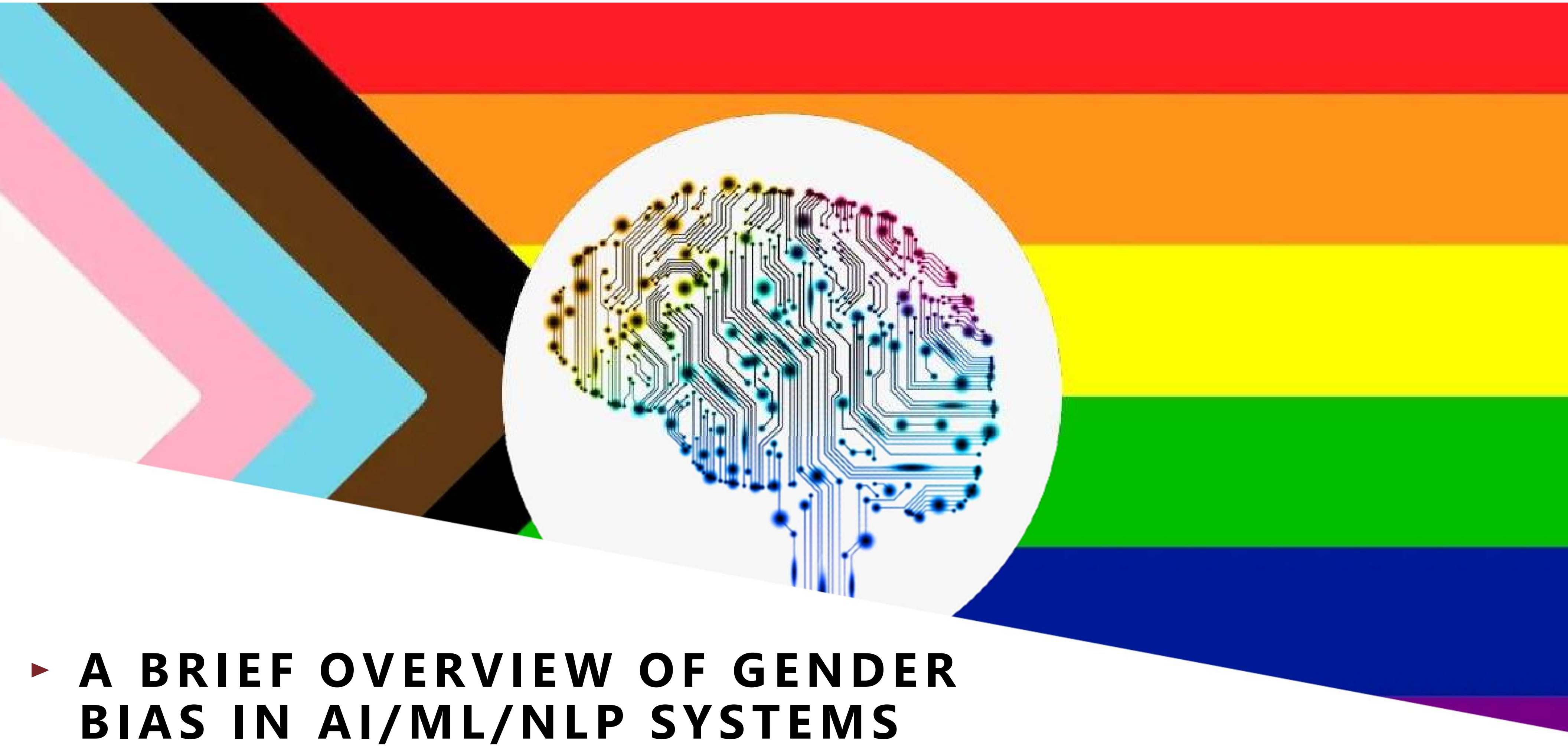


► A BRIEF OVERVIEW OF GENDER BIAS IN AI/ML/NLP SYSTEMS

CHRISTOPHER E. RASHIDIAN, M.S.D.A.

JUNE 27, 2022

RCODI
open.digital.innovation

▸ OUTLINE

1 Gender Bias in Artificial Intelligence, Machine Learning, and Natural Language Processing Models

2 Identifying and Mitigating Gender Bias in Models

3 References

GENDER BIAS IN AI, ML, & NLP MODELS

APPLICATION OF MODELS IN DAILY LIFE

Modern society relies on machines to automatically analyze natural languages:

- Personal assistants like Siri;
- Machine translation systems like Google Translate; and
- Others such as those that track job applicants.

▸ GENDER BIAS IN AI, ML, & NLP MODELS

GENDER BIAS

Could be defined as dominance of one gender over another – occupations or social roles:

- This is based on historical roles in that men were more likely to engage in tasks requiring strength and speed and women have been more likely to stay home and engage in family tasks;
- Less dominant gender is underrepresented and stereotypes appear (i.e. nurses tend to be females and doctors tend to be males); and
- This is a two-fold system:
  - (i) Could be a tool to identify gender bias in social context.
  - (ii) It produces gender biased system.

LINGUISTICAL APPROACH IN GENDER BIAS – TAXONOMY OF GENDERED PRONOUNS

Generic pronouns are those that do not identify sex due to them being generic as they discuss humans in general. Examples include:

1. He, his, him – A good student knows that he should always study to earn high marks;
2. She, hers, her – A flight attendant should ensure that she takes her break; and
3. Man – A good pilot should always man the cockpit.

▸ GENDER BIAS IN AI, ML, & NLP MODELS

LINGUISTICAL APPROACH IN GENDER BIAS – TAXONOMY OF GENDERED PRONOUNS

Additional bias has been identified in sexism, which includes:

1. Hostile Bias – Where men are seen as more powerful than women and women are a threat to men's dominance. An example of such bias is *women always get more upset than men.*
2. Benevolent Bias – Where male dominance is asserted in a more chivalrous tone while expressing their affection for women in return for their acceptance into limited gender roles. An example of such bias is *I am surprised at how knowledgeable in physics you are for a girl.*
3. Occupational Bias – This is derived in gender-typical social roles and reflects the sexual division of labor and gender hierarchy. An example of such bias is *Pilots are men and flight attendants are women.*
4. Exclusionary Bias is (i) the explicit marking of sex in unknown gender-neutral entities (i.e. *mankind, brotherhood*); (ii) gender-based neoglism in where newly coined words and expressions are in the process of being adopted into mainstream culture (i.e. *manscape, chick flick*), and (iii) gendered word ordering where the masculine version of the word proceeds the feminine version of the word (i.e. *Mr. and Mrs., boys and girls*).

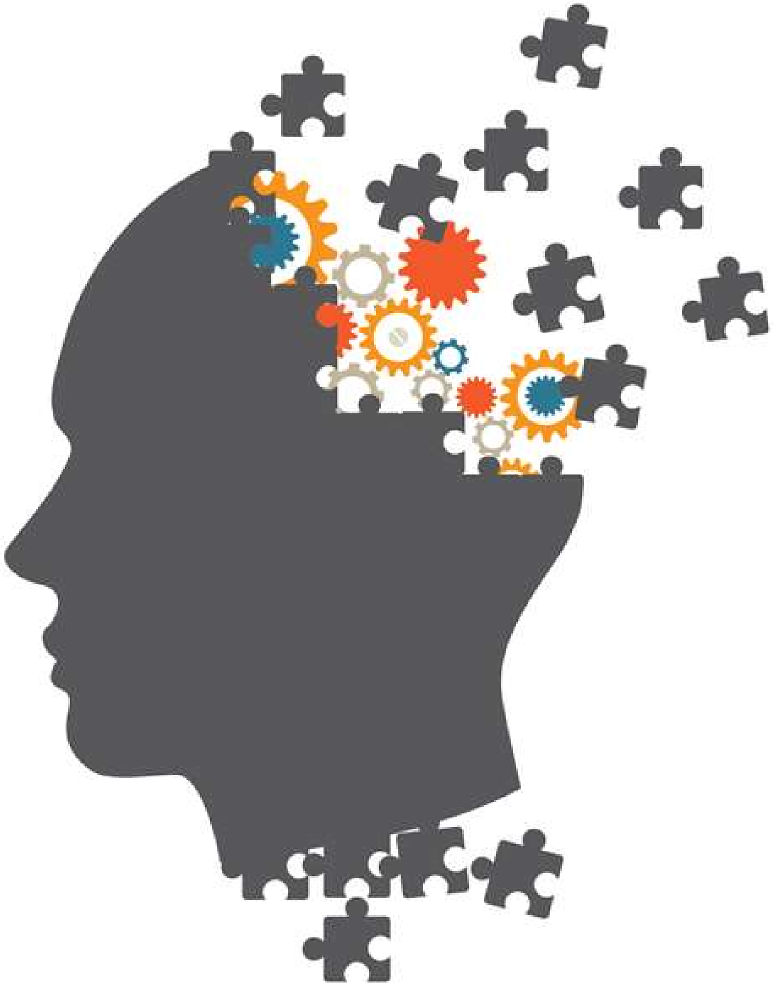
IDENTIFYING AND MITIGATING GENDER BIAS IN MODELS

CAUSES OF GENDER BIAS IN AI, ML, AND NLP

Causes of bias in artificial intelligence, machine learning, and natural language processing could be:

1. Incomplete or skewed data sets;
2. Labels used in training; and
3. Features and modeling techniques, including:
    a) Analogies
    b) Classification
    c) Clustering

These techniques are scrutinized because they do not focus on the impact of real-world applications.

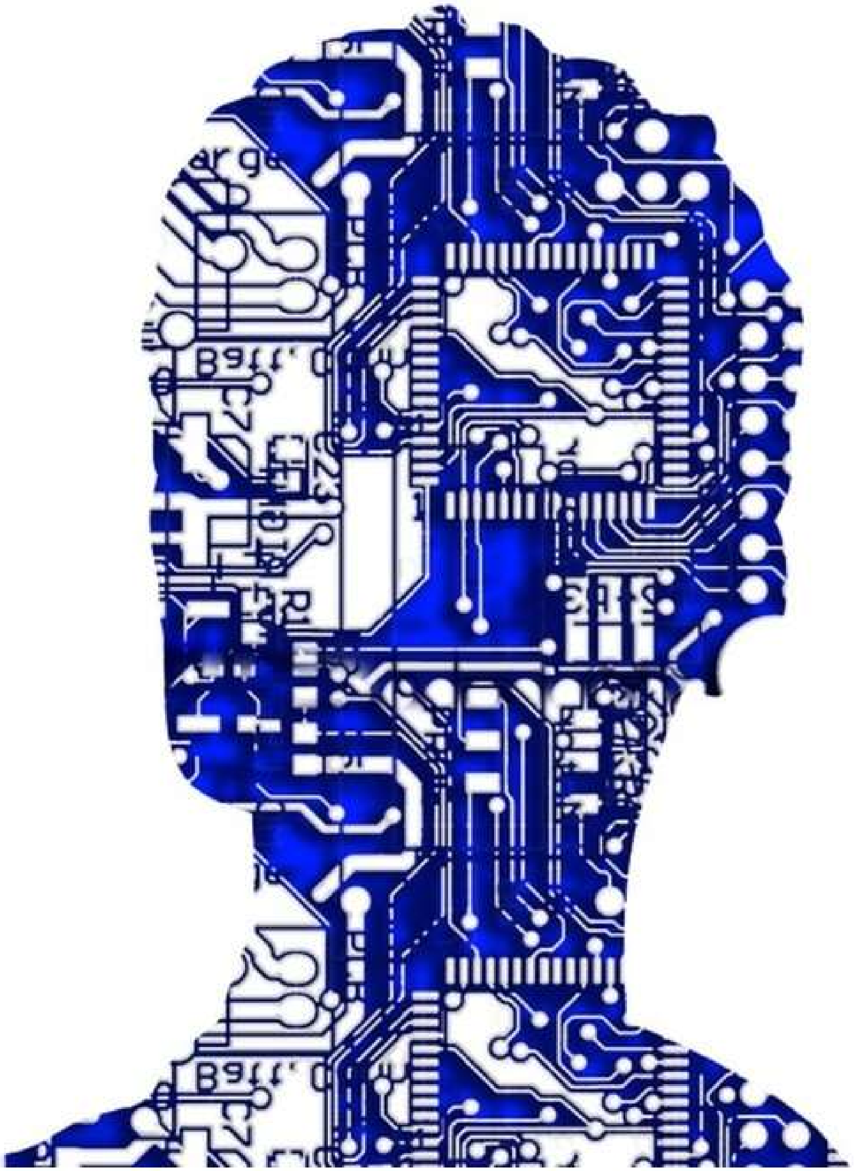


▸ IDENTIFYING AND MITIGATING GENDER BIAS IN MODELS

OBSERVING AND IDENTIFYING GENDER BIAS IN AI, ML, AND NLP

The three ways to observe and identify bias in artificial intelligence, machine learning, a natural language processing are

- Adopting psychological tests;
- Analyzing gender subspace in embeddings; and
- Measuring performance differences across genders.



▸ IDENTIFYING AND MITIGATING GENDER BIAS IN MODELS

ORGANIZATIONAL EFFORTS TO MITIGATION GENDER BIAS IN AI, ML, AND NLP

One way to mitigate gender bias in artificial intelligence, machine learning, and natural language processing is a change in company culture, including:

- Ensure diversity is present in training samples;
- Ensure that staff labeling datasets come from diverse backgrounds;
- Encourage staff to test accuracy for different demographic categories to identify bias; and
- Collect more training data for sensitive groups and apply debiasing techniques.

DEBIASING TECHNIQUES AND METHODS

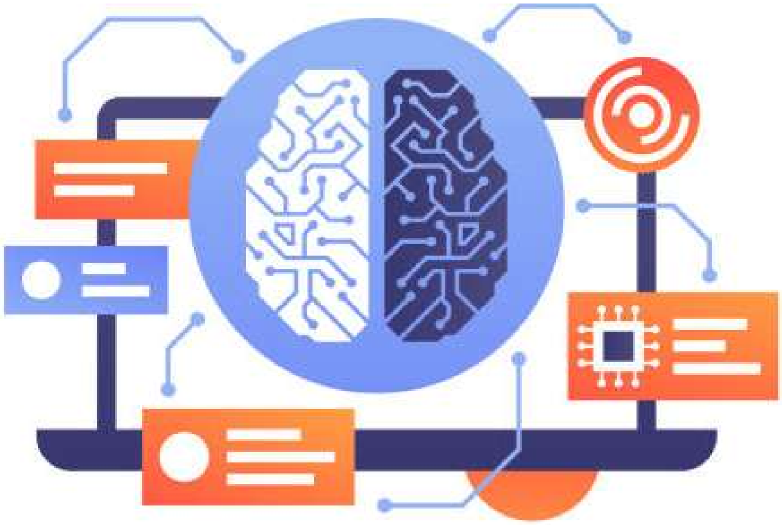
One method of debiasing the model involves the manipulation of data which includes:

- Data augmentation;
- Gender tagging;
- Fine-tuning bias; and
- Debiasing gender in word embeddings, including (i) removing gender subspace, and (ii) learning gender-neutral word embeddings.

DEBIASING TECHNIQUES AND METHODS

The second method of debiasing the model involves the adjustment of algorithms that are used, including:

- Constraining predictions; and
- Adversarial learning, which is adjusting the discriminator.



COLLABORATION ACROSS DISCIPLINES

This is not a computer programming issue, but this should be approached through an interdisciplinary lens by marrying social sciences and S.T.E.M.

REFERENCES

Costa-jussà, M. R. (2019). An Analysis of Gender Bias Studies in Natural Language Processing. *Nature Machine Intelligence*, *1*(11), 495–496. https://doi.org/10.1038/s42256-019-0105-5

Doughman, J., Khreich, W., El Gharib, M., Wiss, M., & Berjawi, Z. (2021). Gender Bias in Text: Origin, Taxonomy, and Implications. *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*. https://doi.org/10.18653/v1/2021.gebnlp-1.5

Feast, J. (2020, October 8). *4 Ways to Address Gender Bias in AI*. Harvard Business Review. Retrieved June 19, 2022, from https://hbr.org/2019/11/4-ways-to-address-gender-bias-in-ai

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2020). Gender Bias in Chatbot Designs. *Chatbot Research and Design*, 79–93. https://doi.org/10.1007/978-3-030-39540-7_6

Leavy, S. (2018). Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning. *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*. https://doi.org/10.1145/3195570.3195580

Leavy, S., Meaney, G., Wade, K., & Greene, D. (2020). Mitigating Gender Bias in Machine Learning Data Sets. *Communications in Computer and Information Science*, 12–26. https://doi.org/10.1007/978-3-030-52485-2_2

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/p19-1159

CR

QUESTIONS