Fairness in Artificial Intelligence Research Brief

Lab 58 Technology Research Brief

June 2022

A multitude of industries increasingly make use of artificial intelligence (AI) machines, computer systems that can perform tasks without the requirement of human intelligence. AI machines can uncover criminal activity and solve crimes, and identification using facial recognition technology is used as often as fingerprinting.[1] However, models of AI machines, although extremely efficient and productive, can be unfair and have machine learning bias. Bias is a prejudice for or against one person or group, typically in a way considered to be unfair. Machine learning bias is defined as an error from incorrect assumptions in the algorithm or systemic prediction errors that arise from the distribution properties of the data used to train the AI model. Thus, the AI model consistently makes the same mistakes related to certain groups of individuals. These errors can have immense social and economic consequences.

Although existing bias in society can be embedded in AI systems, detecting these undesired correlations between some features can allow users to mitigate and avoid them. Fairness in AI refers to the attempts at correcting program bias in AI systems. Fairness in AI works as a preventative and reservable measure to combat AI bias. AI developers must identify and suggest a comprehensive view for Fairness in AI technologies, one that covers both the technical and nontechnical dimensions of AI fairness to build the most effective framework, techniques, and policies. This brief highlights an example of implicit bias in AI machines related to sex and skin tone, consequences of implicit bias in AI, techniques for Fairness in AI, and use cases for Fairness in AI applications.

KEY TAKEAWAYS

Fairness in AI works to decode or reduce integrated algorithmic biases by debiasing the AI dataset and model.

Use cases for Fairness in AI include reducing or eliminating subconscious human biases, such as biases related to sex, skin color, and sexual orientation—as well as biases in natural language processing patterns.

Fairness in AI works to prevent allocation harms, quality of service harms, and representation harms.

How AI Systems Can Be Biased

AI machines can incorporate unconscious bias in many ways. For example, some AI machines have been shown to reflect human biases related to gender and skin color. MIT conducted a research study to determine whether there was bias in AI machines that contained data on darker- and lighter-skinned males and females. (Classification is from the study source; individuals outside of the male/female binary were not analyzed in this study.) The MIT research study found bias against females with dark skin in three general-purpose facial-analysis systems.[2] The study found that across the three systems, "the error rates for gender classification were consistently higher for females than they were for males, and for darker-skinned subjects than for lighter-skinned subjects."[2]

[1] Marr, B. (n.d.). *What is the impact of artificial intelligence (AI) on society?* Bernard Marr & Co Future Business Success.

[2] Hardesty, L. (2018, February 11). Study finds gender and skin-tone bias in commercial artificial intelligence systems. *MIT News.*

MIT Media Lab published this video, which depicts the classification of four sex/skin-tone groups and three facial-analysis systems. All three facial-analysis systems had highly accurate detection rates for the light-skinned male, light-skinned female, and dark-skinned male groups. However, all three systems had a dramatically low accuracy detection rate for the dark-skinned female group. The video also shows the large error difference between the dark-skinned female group and light-skinned male group in IBM's facial-analysis systems.

This means that dark-skinned females were 34.4% more likely to be misclassified than light-skinned males. In the MIT research study, photos of dark-skinned individuals of the female sex were frequently not recognized as featuring a human face, or these individuals were misclassified by gender.[2] AI machine implicit bias can cause allocation, quality of service, and representation harms for the discriminated population group.

Steps Toward Fairness in AI

Setting up a comprehensive framework around Fairness in AI involves both technical and nontechnical measures. Observing the following steps will construct AI systems with fairer final outcomes and decisions.

Nontechnical

1. **Define a set of AI principles to guide the organization.** Every AI technology brought on by a company should abide by a set of principles constructed around the advancement of AI ethics. Multiple dimensions of ethics should be considered as defining principles, including fairness, explainability, robustness, privacy, and transparency.[3]
2. **Build the team, training, and governance model to operationalize those principles.** To successfully abide by the ethical principles set up by the organization, a strong team, educational training, and governance model need to be established. AI developers and designers are often not aware of bias in their models and do not have the knowledge to identify what is fair and appropriate for a given scenario.[3] To educate the AI team, IBM suggests training devoted to "education and awareness initiatives for designers, developers, and managers."[3] Additionally, the governance model should focus on diversifying the AI team composition and establishing multiple stakeholder relationships.[4]

Technical

3. **Identify and mitigate bias in AI predictions.** From a technical standpoint, a company must build tools and strategies that uphold its Fairness in AI principles. These Fairness in AI strategies should be used to detect and mitigate the presence of AI machine bias.[4] Checking and mitigating both data bias and model bias can support Fairness in AI. Figure 1 shows the use of these techniques during the AI developmental process.
   a. Solution 1: *Data Bias Checking and Mitigation*
   A preprocessing technique for Fairness in AI is to debias the dataset before inputting data into the AI model. This technique removes the information correlated to sensitive attributes while preserving information.[5] It is best to use this technique during the data collection phase (Figure 1). Debiasing the dataset can also be a postprocessing technique, removing datasets that contain bias and editing data in a way that makes the AI output fairer.[5] This technique is used during the feedback phases of AI development.

   Debias a Dataset

      i. Sample size equality: Sufficient coverage of all groups represented in data. Do not cover one group more than another.
      ii. Diversify the dataset to be representative of the population with which the AI machine will be working.
      iii. Remove skewed datasets.

[3] Rossi, F. (2020, November 5). How IBM is working toward a fairer AI. *Harvard Business Review.*
[4] Chatila, R., & Rossi, F. (2021, January 22). AI fairness is an economic and social imperative. Here's how to address it. *World Economic Forum.*
[5] Zhong, Z. (2018, October 21). A tutorial on fairness in machine learning. *Towards Data Science.*

b. Solution 2: *Model Bias Checking and Mitigation*
A testing-time technique for Fairness in AI is to debias the model by reevaluating the model's labels, features, and tasks. This technique works by adding "a constraint or a regularization term to the existing optimization objective."[5] These constraints force the model to improve fairness by keeping the same rates of measures for the protected group and other individuals. It is best to use this technique during the Test and Implement phases of AI development (Figure 1).

Debias an AI Model

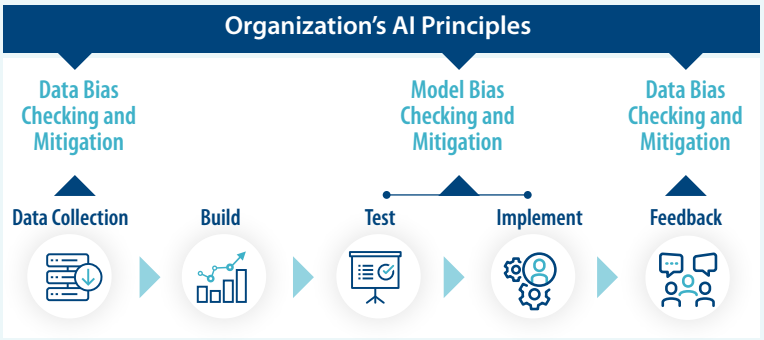
i. Remove proxies of sensitive attribute.

ii. Add reliable and informative features for all groups (to avoid noise for one group over another).

iii. Recode trained models of tainted examples.

4. **Invest in quantitative tools to help reduce AI bias.** In addition to building tools and data that detect and mitigate bias, there should be an investment in established Fairness in AI tools. For example, IBM's *AI Fairness 360* is "an open-source library to detect and mitigate biases in unsupervised learning algorithms."[6] The library "enables AI programmers to test biases in AI models and datasets with a complete set of metrics" and "mitigate biases with packaged algorithms."[6] Examples of tools within the *AI Fairness 360* library include Learning Fair Representations, Reject Option Classification, and Disparate Impact Remover.[6] Investing in quantitative Fairness in AI tools that are already trained and tested will build collaboration and standardization of AI fairness. In doing so, AI developers will be incentivized to work together and improve on existing AI models, rather than produce "hundreds of mediocre tools."[7]

Figure 1: IBM AI Fairness 360



Source: Bayern, M. (2020, June 4). IBM AI Fairness 360 open-source toolkit adds new functionalities. *TechRepublic*.

Use Cases of Fairness in AI

Facial-Analysis Systems

Facial analysis is an erupting industry for AI systems. Companies may use facial-analysis AI to identify a person's gender, recognize a criminal suspect, unlock a person's phone, and more. However, facial-analysis systems have been proven to have gender and skin-tone bias against certain groups, as discussed in the previously mentioned MIT Media Lab video.

Fairness in AI tools can be used to prevent the skin-color and gender bias in systems using facial recognition. For example, Joy Buolamwini, a participant in the MIT study, proposed the Fairness in AI strategy of varying the racial and gender representation in image datasets. Buolamwini proposed that datasets should include a wider variety of features, such as eye shape and width of eyebrows, to allow AI to more accurately identify sex regardless of the subject's skin tone. This Fairness in AI strategy increased the facial-analysis AI model's accuracy in identifying a person's sex, regardless of skin tone, and continues to progressively work against discrimination of any demographic group.[8]

Natural Language Processing

Natural language processing has been established as one of the biggest industries for AI systems. However, although AI machine learning models achieve high performance on many language-understanding tasks, these models lack the tools to reduce group-implicit biases.[9] An example of model language bias was OpenAI's GPT-2 model, trained simply to predict the next word in 40GB of internet text. This model was found to link specific groups with negative connotations (such as linking the word "Muslim" to violence).[10]

Fairness in AI tools can be used to prevent the penalization of gendered words and language stereotypes in AI systems that analyze words and phrases to make decisions. For example, after application of the Hard Debias algorithm and Double-Hard Debias algorithm tools, AI machines have been more effective at identifying characteristics associated with "female" and "male" without negative stereotypes.[9] According to Jerry Wei, "The consequences of letting biased models enter real-world settings are steep, and the good news is that research on ways to address NLP [natural language processing] bias is increasing rapidly."[9]

[6] Dilmegani, C. (2022, September 12). Bias in AI: What it is, types, examples, & 6 ways to fix it in 2022. *AI Multiple*.
[7] Heaven, W. D. (2021, July 30). Hundreds of AI tools have been built to catch covid. None of them helped *MIT Technology Review*.
[8] Buolamwini, J. (2019, February 7). Artificial intelligence has a problem with gender and racial bias. Here's how to solve it. *Time*.
[9] Wei, J. (2020, September 1). Bias in natural language processing (NLP): A dangerous but fixable problem. *Towards Data Science*.
[10] Ngo, K. (2021, February 24). Bias in large language models: GPT-2 as a case study. *Ethical Legal Data Science*.

Medical Field

AI machines have been used in the medical field to help diagnose and treat medical conditions. These AI machines were trained to help doctors evaluate what they are seeing and to make important medical decisions faster and simpler. Although these AI machines were designed to aid the medical industry, some models became potentially harmful. For example, recent AI tools used to detect COVID-19 were found to make decisions based on the most common feature among pictures of COVID-19 patients in their dataset, such as "children" or "a patient that is laying down." Thus, the AI learned to identify "children" or "a patient who is lying down" as COVID-19, instead of identifying the actual infected COVID-19 patients.[7]

Fairness in AI tools can be used to combat potentially harmful outcomes in AI medical machines. AI teams need to collaborate with clinicians and researchers. Researchers need to share their models and disclose how they were trained so others can test those models and build on them.[7] For example, the World Health Organization is considering an emergency data-sharing contract during international health crises. This contract would let researchers move data across borders more easily and establish "data readiness" for the AI medical industry.[7] The sharing of and collaboration on AI data and models would improve AI accuracy and generalizability.

Technical Limitations

Although Fairness in AI tools help mitigate potential bias in AI machines, AI systems will never be fully accurate. Furthermore, all AI machines are built by humans, and humans all have implicit bias, whether conscious or not. The reproduction of human implicit bias will always be a consequence of machines developed by human intelligence.

Future Trajectory of Fairness in AI

The future trajectory of Fairness in AI is bright. Now that AI is providing an increasing number of recommendations to human decision makers, machine fairness is critically important. AI development is increasingly addressing the ethical concerns of privacy, bias and discrimination, and human judgment. Developers are directly encoding additional safeguards within AI algorithms to reduce human biases. Current and new Fairness in AI technologies and strategies can combat AI machine biases. Developers must revisit Fairness in AI tools often to stay up to date with the current predictors of fairness. Additionally, Fairness in AI tools will need to be improved and scaled as AI continues to progress.

RTI and Fairness in AI

The consequences of implicit AI bias entering real-world settings are severe. Bias is a critical issue that must be addressed during the development of every AI machine. Thus, a company that values equity and fairness should not deliver AI/machine learning research and development without explicitly addressing Fairness in AI. With the commitment to Fairness in AI, AI developers can mitigate the implicit biases found in AI and develop machines that make the fairest decisions. Constructing Fairness in AI technologies and supporting Fairness in AI principles will elevate the respect of human values and prevent unnecessary harms to marginalized demographic groups.

Work With Lab 58

Thanks for your interest in our work! Our researchers and developers are actively exploring use cases for and tools to increase Fairness in AI, and we want to help you explore opportunities to work with the technology.

Please email us at Lab58@rti.org. We will set up a 30-minute, one-on-one chat to discuss opportunities and answer any questions. We are interested in partnering with you to find a solution that meets your needs.

For more information, contact Lab58@rti.org.