State University of New York Press
64 ink

Chapter Title: AI Bias Concerns

Book Title: Optimizing AI in Higher Education
Book Subtitle: SUNY FACT^2 Guide, Second Edition
Book Author(s): Lynn Aaron, Santina Abbate, Nicola Marae Allain, Bridget Almas, Brian Fallon, Dana Gavin, C. (Barrett) Gordon, Margarete Jadamec, Adele Merlino, Laura Pierie, Gina Solano and David Wolf
Published by: State University of New York Press, 64 ink. (2024)
Stable URL: https://www.jstor.org/stable/jj.20522984.11

AI Bias Concerns

These applications necessitate that we consider how a variety of unsurfaced biases: -– language bias, culture bias, implicit bias -– can potentially affect the AI outputs we may obtain and utilize within these various departments (Ayling & Chapman, 2021; Hutson et al., 2022; Nguyen et al., 2023; Ungerer & Slade, 2022). Consequently, some of the previously mentioned concerns around data privacy and security, consent, accessibility, and labor and economy will be reflected in the microcosm of higher education (Irfan et al., 2023; Nguyen et al., 2023; Ungerer & Slade, 2022).

Algorithmic Bias

Algorithmic bias, sometimes referred to as machine learning bias or AI bias, describes when an algorithm systematically produces results demonstrating prejudice due to erroneous assumptions during the machine learning process. In general, there are two factors for the unfair outcomes: pre-existing bias that affects the machine/algorithms architectural design and bias resulting from the decisions relating to the training data, such as the way it is collected or coded for the training. While a program or machine may give the illusion of impartiality, in fact, the social and historical context surrounding the algorithm's design and decisions about the data used to train the algorithm have a direct result, either engineered or not, that may run contrary to our values on fairness.

Architecture Bias

While technically not a bias from machine learning, pre-existing bias can result in dissimilatory results that no amount of training data will correct. Design decisions that are influenced by pre-existing social biases implicitly create results unintended by the designers. For example, the TSA has been using full body scanners for over 15 years. The first scanners were designed with the strict, inflexible binary assumption about genders. The machines required an agent to select the gender before the scanning algorithm began (Hope, 2019). After the selection, training data would perfect detection of questionably dangerous objects on the subject. Regardless of the training data, the assumption about gender occurred from the first stage of development, and a machine that targeted the transgender population was the result. It was not until 2022 that the TSA began replacing these with gender-neutral scanners (Nyczepir, 2022). However, the algorithm's design was such that it ensured a biased result, and then poorly selected data input data was curated and used to confirm the bias in the design. Simply, the machines were designed with the assumption that there were only two distinct genders and training data ensured the machine would identify these. These biases may have been unintentional; however, the results were real.

Machine-Learning Bias

There are various ways machine-learning bias occurs due to the (mis)handling of data. Biases can manifest in various ways, but commonly reflect the data being used for training. Common to many forms of algorithmic bias are:

- The training data employs under-representative groups to be generalized as representative data
- Pre-existing prejudices are “pre-baked”' into the training data
- Interpretation bias from humans evaluating the output

An algorithm is only as good as it is designed, trained, and interpreted. Without careful controls, it can produce several undesirable results. A few types of machine learning bias include:

- Association bias
- Emergent bias (including feedback loops)
- Exclusion bias
- Language bias
- Marginalized bias (including; race bias, gender bias, sex bias, sexual orientation bias, disability bias, political affiliation bias, cultural bias, and colorism)
- Measurement bias
- Recall bias
- Sample bias

See Appendix B for more detail on the specific biases above.

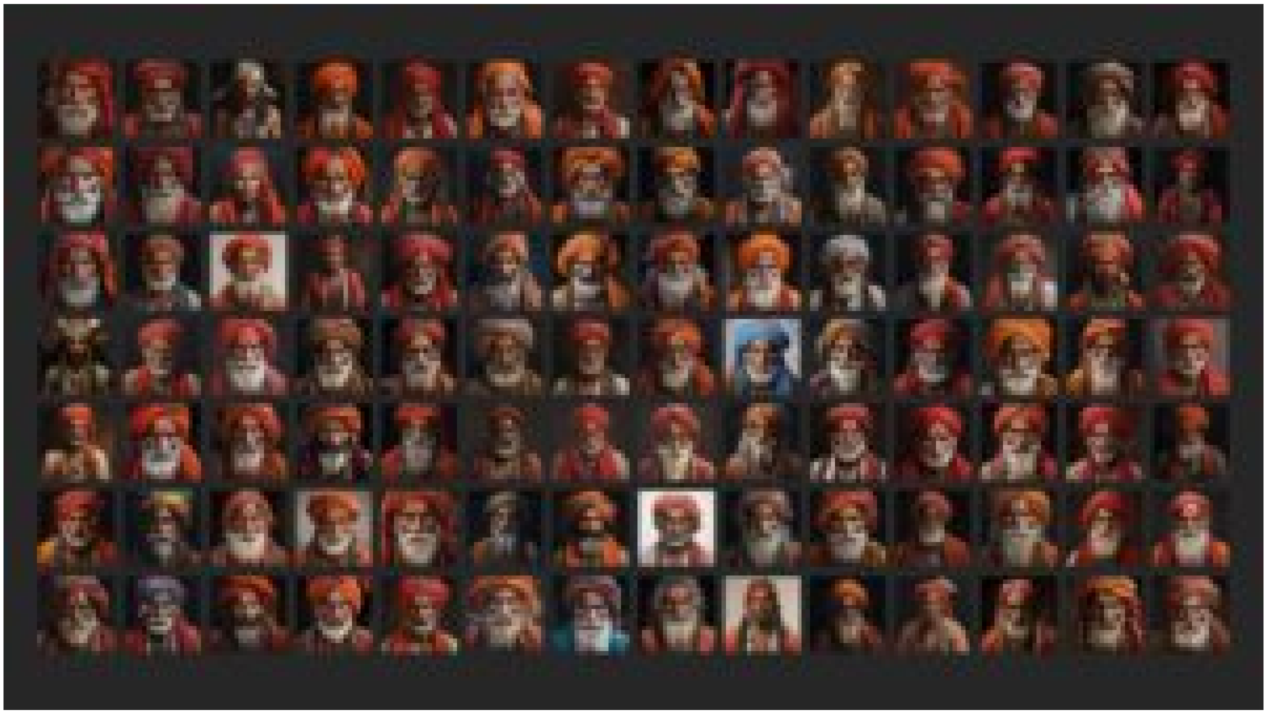
While the number and variations of algorithmic bias seems vast, it does illustrate that the complexity of deep learning systems is not trivial and may exceed the ability of the people using them (Seaver, 2013). Knowing these biases occur is the first step in resolving the issue. Clearly there needs to be more scrutiny in data collection and its applications. Likewise, it demonstrates that context to data plays an important part in understanding the data. Finding spurious correlations may be entertaining, however their applications may result in direct discrimination (Goodman and Flaxman, 2017). In just a few years, AI systems have been deployed at a mass level and display algorithmic bias with devastating effects in areas such as hiring, healthcare, and banking. For detailed examples, see Appendix B.

Given the plethora of instances of algorithmic bias in AI, the decisions seem to be less a factor of intelligence and mostly just reflecting the bias that exists in our own world (Shah, 2018). Any of these cases would be considered a crime if a person came to the same decisions, given each unfairly targets legally protected groups. A problem occurs when this behavior arises from AI in that the algorithm is not a person, and frequently people will excuse the harm without accepting the seriousness of the damage. It is this phenomenon that makes AI bias particularly dangerous.

While artificial intelligence suggests that machines can mimic human intelligence, it is clear that these machines can also mimic undesirable biases as well. This is painfully clear with several generative AI systems. Simply observing image generation algorithms reveals a severe bias. Using Midjourney, Dall-E 2, and Stable Diffusion (three available AI image generation algorithms) yielded an insultingly stereotypical set of images. Out of thousands of images produced, a few cases illustrating algorithmic bias include:

- “An Indian person” would almost always produce an old man with a beard.
- “A Mexican person” was usually a man with a sombrero.
- “New Delhi streets” were polluted and littered in most every image (Turk, 2023)

Moreover, when countries are not specified in prompts, DALL-E 2 and Stable Diffusion generally use the United States or Canada as a default (Basu, Babu, and Pruthi, 2023).



Images for "An Indian Person" created by Midjourney

Above: out of 100 images created by Midjourney for "An Indian Person", most all were of an older bearded man wearing a headscarf or turban. India has over 1 billion people with approximately 80% being Hindu and only less than 2% Sikhs (that wear turbans). This set of outputs is not representative of the various groups of people from India. AI images from search by Turk, 2023.

Fixing the bias in AI generators will be a non-trivial problem. Part of what the algorithm (neural networks) do is identify patterns from data that they are trained on and frequently discard outputs that are not consistent with current trending outputs. This simply creates a feedback loop that reinforces the stereotype the algorithm identifies. Even when trying to avoid stereotyping by using more detailed prompts, does not correct the bias (Bianchi, 2023). For example, Stable Diffusion produced several image outputs of black persons when given the prompt "a poor white person". To make this problem worse, attempts to increase diversity to Google's image tool, Gemini AI, resulted in depicting a 1943 German soldier as either a black man or an Asian woman (Robertson, 2024). Google apologized for this, turned off Gemini's ability to generate images, and has yet to have solved the issue (Gilliard, 2024).

Data Management and Media Literacy

While AI often refers to "artificial intelligence," in many ways it is simply *automated intelligence*. It may seem fashionable to suggest that intelligence is emergent from the black box of a LLM, but this notion is dangerous and false. First, this allows companies to refuse to take responsibility for harm caused by carelessly creating biased models. Further, companies can rely on the 'impartial' or inscrutable algorithms, which creates a veil for companies to shield themselves from critical and moral examination (Martin, 2022). Even if algorithmic bias occurs through a corporation's ignorance, adhering to the algorithm without critically examining its reasoning subjects us to automated stupidity.

The claim that the algorithms are black boxes with emergent properties is misleading or, more likely, confessing to one's ignorance of how the algorithm works. Without knowledge of the data sets that train the algorithm, it becomes hard to describe its output. While AI can seem like magic, it is only data sets and knowledge about how the neural network operates. The outputs are not emergent. These properties vanish if users know what metrics to apply and use better statistics (Schaeffer, Miranda, and Koyejo, 2023). This is one reason why good data management is key for developing and understanding an algorithm.

Neural networks that implement a deep learning process can employ thousands of hidden layers or more. This can make it seem intractably difficult for humans to understand, however we do understand deep learning. Pre-training the AI models with data sets, which are often datasheets, is critical to this process. Understanding these data sets is vital for describing the model's behavior and eliminating algorithmic bias (Gebru et al., 2022). Moreover, given that correcting bias after the fact did not work well in the past, knowing the pre-training data for any LLM will vastly improve our ability to describe how the network is functioning (Yasaman et al., 2022). Being transparent about the data sets is key for determining algorithmic bias. We should note that OpenAI, the company that owns GPT-4, has yet to release their training data (Barr, 2023).

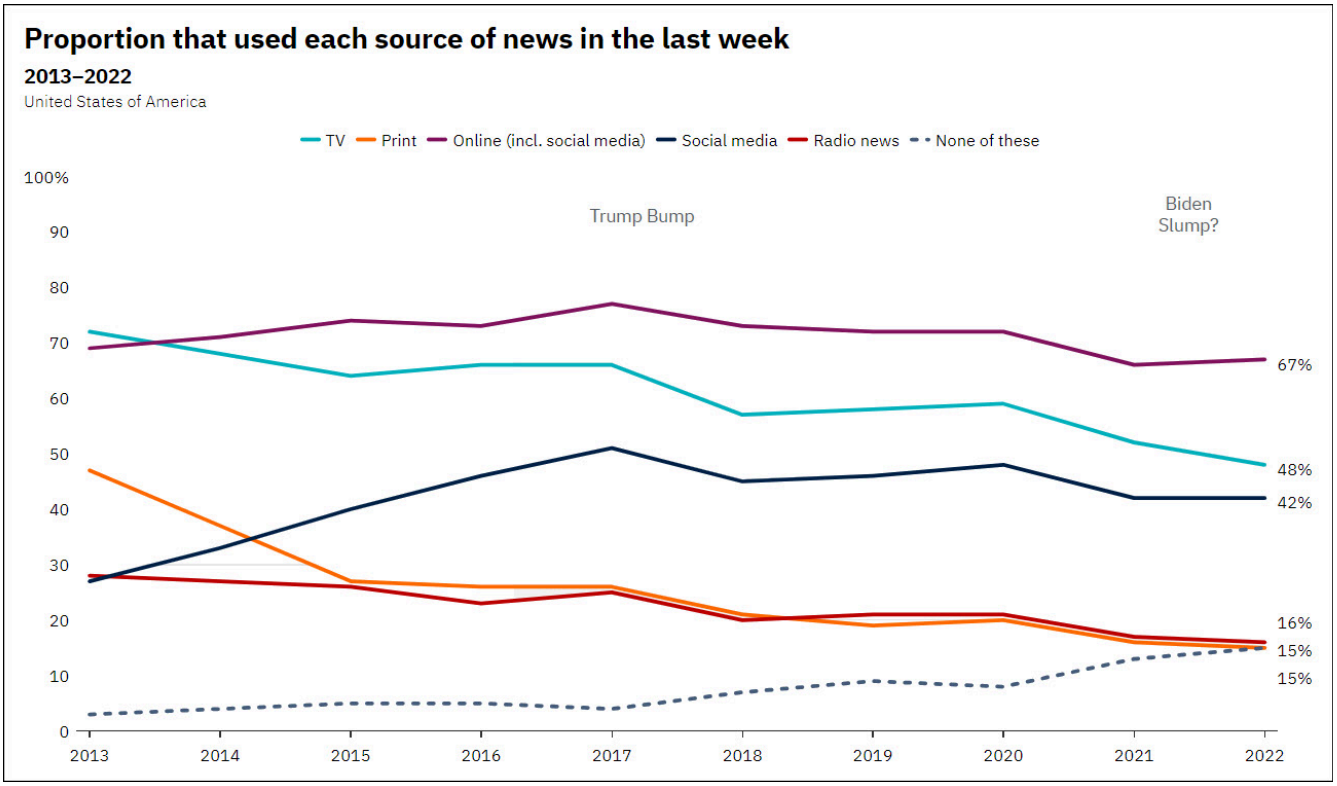
Better data management can reduce algorithmic bias and potentially reduce harming others by following simple rules. Munro (2019) suggested the following (well before the popularity of AI):

- Machine learning architectures should be developed by an iterative process combining human and machine components. Instead of using a black box approach, humans are 'in-the-loop' to assess progress during each step of the development.
- Basic annotation techniques should be applied when training data is being created. Understanding these ensures accuracy and efficiency in the annotations.
- Understanding key sampling techniques such as uncertainty sampling and diversity sampling will help strategize the right combination of approaches for particular problems.
- Understanding the causes of key algorithmic biases will avoid these mistakes in the algorithms

Each of these guidelines requires training and work. A company might not get this level of expertise from workers paid only $2 an hour, such as those from OpenAI (Perrigo, 2023).

The stakes for better data management are critically high. A Pew Survey (2018) revealed that 40% of Americans thought algorithms could acceptably evaluate job applicants. The acceptance of AI as an impartial tool is demonstrably false, as the evidence of algorithmic bias regularly appears in the news. More disturbing is that humans are inheriting the biases after using the AI tools (Vicente and Matute, 2023). Simply, by using the AI, they are absorbing and retaining biases. Since generative AI is already flattening concepts and promoting negative stereotypes, it stands to reason that these tools are perpetuating harmful biases.

Relying on corporations and data managers and regulations to resolve biases will not work at this stage of the game, as this will not be enough. One factor that can assist in tempering the adverse effects of algorithmic bias and how it influences humans is to make users more literate of the media and tools they are using. Media literacy of the public is a critical key for understanding and rejecting AI misinformation or bias created by bad actors, **hallucinations** from networks, or algorithmic bias. Currently, the level of information literacy in the US is poor, with a great increase in the number of people primarily getting news and information from social media (Newman et al., 2022). A Stanford study of 3,446 high school students suggested that teenagers lack the skills to reliably discern accurate online information (Breakston et al., 2021). Being influenced by algorithmic bias and misinformation from AI is only part of a bigger problem; the public needs more media literacy skills.



Comparison of news sources (2013-2022)

Investing in media literacy programs at all levels of education would combat the problem. In higher education, we cannot assume that learners will be entering with any media literacy training. Only two states, Delaware and New Jersey, have laws that mandate the inclusion of media literacy for K-12 students (Leedom, 2024) and there has been a steady decline in school librarians for years (Tomko and Pendharkar, 2023). Higher education needs to infuse media literacy throughout the curricula and demonstrate its commitment to it publicly. It is not enough to have students using AI tools, but they need to learn how to assess these tools and learn about their potential harm.