Title

Who Trains the Machine? Standpoint Theory and the Social Origins of Bias in Artificial Intelligence

Author

Turner, Ademole

Publication Date

2026-03-16

Who Trains the Machine? Standpoint Theory and the Social Origins of Bias in Artificial Intelligence

Ademole Turner[1]

[1]Chemical and Biomolecular Engineering Department, University of California, Los Angeles

*ade2x@ucla.edu

Abstract
Artificial intelligence systems are often treated as neutral technologies. Using standpoint theory, this paper examines how AI training datasets and model design embed and reproduce racial and gender biases present in existing social structures.

INTRODUCTION
Artificial intelligence (AI) technologies are increasingly integrated into areas such as healthcare, research, digital media, and automated decision making. Machine learning systems now influence hiring tools, recommendation algorithms, predictive policing systems, and medical diagnostics. Because these technologies rely on large datasets and statistical modeling, they are often presented as objective and data-driven solutions to complex problems. However, recent research has demonstrated that AI systems can reproduce racialized and gendered assumptions embedded in their training data. Rather than eliminating human bias, algorithmic systems often reproduce and amplify the inequalities present in the datasets used to train them.

AI models learn patterns from large collections of digital information gathered from the internet, including images, written text, and metadata associated with online content. These datasets reflect social patterns already present in society, including stereotypes, unequal representation, and historical inequalities. When models learn these patterns, they may treat them as statistically meaningful relationships rather than social artifacts. As a result, AI systems can generate outputs that reinforce existing social hierarchies. For example, research on generative image models has shown that prompts describing certain occupations often produce images that associate jobs with specific genders or racial groups [1]. Similarly, studies of computer vision models demonstrate that many systems perform less accurately for individuals with darker skin tones because those populations are underrepresented in training datasets [2].

These patterns raise important questions about how knowledge is produced in technological systems. Artificial intelligence is often framed as a purely technical field focused on optimizing model accuracy and computational efficiency. However, the social context in which AI systems are developed influences how datasets are constructed, what problems researchers prioritize, and how models are evaluated. As a result, AI development cannot be separated from broader social and institutional power structures.

Standpoint theory provides a useful framework for analyzing these dynamics. Developed within feminist epistemology, standpoint theory argues that knowledge is socially situated and shaped by the perspectives of those who produce it. Scientific knowledge is frequently presented as universal and objective, yet standpoint theorists argue that social position influences which questions are

asked, what evidence is considered valid, and how findings are interpreted. Groups that hold institutional power often shape dominant knowledge systems, while marginalized perspectives may be excluded from decision-making processes.

Applying standpoint theory to artificial intelligence highlights the importance of examining who controls dataset construction and model development. Many large-scale AI systems are created by corporations, research laboratories, and technology companies with access to vast computational resources and large datasets. These institutions determine what types of data are collected, how datasets are filtered, and which problems AI systems are designed to address. Consequently, the knowledge embedded in AI systems reflects the priorities and perspectives of these institutions.
This paper examines how AI training data and model design embed racialized and gendered assumptions. By analyzing dataset construction, documented examples of algorithmic bias, and recent attempts to address these issues, the paper argues that AI systems reproduce existing power structures embedded in technological development. Through the lens of standpoint theory, algorithmic bias can be understood not simply as a technical problem but as a reflection of the social structures that shape knowledge production in STEM fields.

This work is in partial fulfillment of the ENGR184 course using the blueprint curriculum in Ref [6,7] and captured in a collection [8].

METHODS

This study applies standpoint theory as a critical framework to analyze the relationship between AI training datasets and biased algorithmic outcomes. Rather than performing computational experiments, the research synthesizes findings from peer-reviewed literature, technical reports, and case studies examining AI bias. The goal is to evaluate how training data and model design influence the behavior of machine learning systems.

Three main areas of investigation are examined. The first involves analyzing the structure of AI training datasets. Large language models and generative AI systems are typically trained using massive datasets composed of web pages, books, code repositories, academic articles, and online images. These datasets are often collected through automated web scraping methods that gather content from publicly available sources across the internet. Once collected, the data are processed through several steps including filtering, deduplication, and labeling to prepare them for training machine learning models [4].

Filtering procedures attempt to remove low-quality or harmful content from the dataset. However, filtering decisions depend on the criteria chosen by developers, which can introduce additional biases into the dataset. For example, some automated filtering systems remove certain types of language or content that may be more common in marginalized communities, unintentionally reducing their representation in the final dataset. Deduplication processes remove repeated documents to prevent models from overfitting to specific sources, while labeling processes assign descriptive tags to images or text so that models can learn relationships between words and visual features.

The second area of investigation examines examples of bias in generative AI systems. Image-generation models such as DALL·E and Stable Diffusion learn to generate images by identifying patterns between visual features and textual descriptions in their training datasets. When users enter prompts, the model produces images based on patterns it learned during training. However, because these models rely heavily on internet data, they often reproduce stereotypes present in online imagery and captions. Documented examples of these biases provide insight into how dataset composition influences AI outputs [1].

The third area of investigation reviews efforts to address algorithmic bias through improved dataset design and evaluation methods. Researchers have developed human-centered datasets that intentionally include diverse demographic representation and obtain informed consent from participants contributing images. These datasets aim to provide more balanced benchmarks for evaluating model performance across different groups [2]. By comparing results from existing models on these datasets, researchers can identify disparities in model accuracy and fairness.

Through this multi-source analysis, the study evaluates how social power structures influence AI development and how standpoint theory can be used to critique dominant assumptions about objectivity in artificial intelligence research.

RESULTS AND INTERPRETATION

Research on generative AI systems provides clear evidence that training data can embed and reproduce social stereotypes. Studies of image-generation models have shown that prompts describing certain occupations frequently produce stereotypical outputs. For instance, prompts asking for images of “housekeepers” often generate pictures depicting people of color, while prompts for “flight attendants” commonly generate images of women [1]. These results reflect statistical relationships present in the training datasets used to build these models.

Additional examples demonstrate how generative AI systems associate geographic and racial identities with particular stereotypes. In some cases, prompts describing wealthy individuals in African contexts generate images depicting rural environments or impoverished conditions. These outputs indicate that the model has learned associations between Africa and poverty from its training data. Because generative models rely on probabilistic pattern recognition, they reproduce the associations most represented in their datasets.

Research on computer vision datasets further demonstrates how dataset composition affects model performance. In studies using diverse human image datasets collected with informed consent, researchers found that many existing AI models performed significantly better for individuals with lighter skin tones and younger age groups. Accuracy declined for darker skin tones and other underrepresented demographic groups [2]. These disparities illustrate how uneven representation within training datasets can produce unequal performance across populations.

The impact of these biases extends beyond image generation and into practical applications. In healthcare settings, machine learning models trained on biased datasets have produced inaccurate

predictions for certain patient populations. Research examining psychiatric prediction models found that algorithms may perform less accurately for Black patients because they fail to account for the social and historical impacts of racism on health outcomes [3]. When datasets fail to capture these contextual factors, the resulting models can misinterpret patterns in the data and produce biased outcomes.

These findings challenge a common assumption in machine learning that larger datasets will naturally reduce bias. Some researchers have argued that increasing dataset size should allow models to learn more diverse patterns and therefore reduce discrimination. However, studies have shown that larger datasets can still reproduce harmful content and stereotypes if the underlying data sources contain biased representations [5]. Simply scaling up data collection does not eliminate the social patterns embedded in the data.

From the perspective of standpoint theory, these results illustrate how AI systems reflect the perspectives of the institutions and developers who create them. Decisions about dataset composition, filtering criteria, and model objectives are made by researchers and organizations with specific social and institutional positions. Because most large AI systems are developed by technology companies and elite research institutions, the perspectives represented in their datasets often reflect dominant cultural viewpoints rather than a broad range of social experiences.

Standpoint theory therefore suggests that addressing AI bias requires more than technical adjustments. Instead, it requires examining the social processes that shape dataset construction and model development. Incorporating perspectives from communities affected by AI systems may reveal biases that remain invisible to developers working within dominant institutional contexts.

CONCLUSIONS

This study demonstrates that artificial intelligence systems are not socially neutral technologies. The datasets used to train AI models and the design decisions guiding those models reflect the perspectives and power structures of the institutions that build them. Through the lens of standpoint theory, algorithmic bias can be understood as a product of the social conditions under which technological knowledge is produced.

Examples from generative image models and computer vision research show that biased datasets can produce unequal outcomes across demographic groups. These findings challenge the assumption that simply increasing dataset size or improving model accuracy will eliminate bias. Instead, addressing algorithmic bias requires careful consideration of how datasets are constructed and who participates in the development of AI systems.

Future work in artificial intelligence research should prioritize transparency in dataset construction, improved documentation practices, and greater participation from diverse communities in AI development. By incorporating multiple perspectives into the design of technological systems,

researchers may be able to identify and reduce biases that would otherwise remain embedded in algorithmic models.

REFERENCES

1. Ananya. *AI image generators often give racist and sexist results: can they be fixed?* Nature, Vol. 627, 28 March 2024.
2. Kalluri, P. R., et al. *Fair human-centric image dataset for ethical AI benchmarking.* Nature (2025).
3. Alenichev, A., et al. *Racism-related stress and artificial intelligence in psychiatric prediction models.* Frontiers in Digital Health (2025).
4. Du, F., Ma, X., Yang, J., et al. *A Survey of LLM Datasets: From Autoregressive Model to AI Chatbot.* Journal of Computer Science and Technology, 2024.
5. Mehrabi, N., et al. *Design AI so that it's fair.* Nature (2018).
6. Lee, E., et al. *Education for a Future in Crisis: Developing a Humanities-Informed STEM Curriculum.* arXiv (2023).
7. Carbajo, S. *Nurturing Deeper Ways of Knowing in Science.* Issues in Science & Technology (2025).
8. Carbajo, S. *Queered Science & Technology Center: Volume 4.* (2026).