MIT Press • Open Encyclopedia of Cognitive Science

Algorithmic Bias

Abeba Birhane[1]

[1]**Trinity College Dublin**

MIT Press

Published on: Mar 16, 2026

URL: https://oecs.mit.edu/pub/b61joemo

Algorithmic bias refers to prejudicial, discriminatory, unjust, inaccurate, or otherwise disparate performance or outcomes from algorithmic systems based on racial, gender, or other attributes of an individual or a group. The concept of algorithmic bias emerged at the intersection of computer science, artificial intelligence (AI) research, critical data studies, human–computer interaction, law, philosophy, and similar disciplines. Although problems and discrepancies at the model level denote the most commonly studied form of bias, the term algorithmic bias is also used as a shorthand to describe a multitude of problems and challenges at various steps of the AI pipeline from ideation, problem framing, training data curation and processing, model training and validation, and deployment as well as emergent issues that arise from interaction with the real world. Potential sources of bias, appropriate metrics to define, measure, and mitigate bias, and the utility and merit of technical approaches to bias mitigation are fiercely debated in the current AI landscape.

History

Early conceptions of algorithmic bias can be traced back to the early days of AI and the philosophy of technology. Having developed one of the first computer systems for the Bank of America in the early 1950s, Joseph Weizenbaum articulated how the computerization of banking allows large institutions to solidify and formalize existing structures and hierarchies that favor the status quo rather than encourage solutions that bring forth structural changes or decentralized solutions. In the 1970s and 1980s, the philosopher of technology Langdon Winner dismantled the idea that technology is a "neutral" tool (Winner, 1978). Winner's work emphasized how those developing technology benefit the most while already disadvantaged individuals and groups pay the highest price, laying the philosophical foundation for subsequent work in algorithmic bias (Winner, 1978).

Much early work laid the foundation for broader issues that arise with automation, including the challenges of automating complex cognitive, psychological, and social phenomena (Winograd & Flores, 1986) as well as the downstream impact of such automation (Winner, 1980). Orthogonal to this were more technical efforts to formalize, quantify, and measure algorithmic bias using metrics such as "fairness index" (Jain et al., 1984). By the late 1990s, algorithmic bias emerged as a fully fledged academic endeavor. Friedman and Nissenbaum (1996) put forward three categories of bias in computer systems: *preexisting* (rooted in social institutions and practices), *technical* (arising from technical constraints), and *emergent* (arising in the context of use). Alongside the explosion of AI and machine learning research marked by the "AI revolution," the late 2000s and early 2010s saw an explosion of academic work on algorithmic bias, including Barocas and Selbst's (2016) systematic analysis of how algorithmic systems embed and perpetuate discrimination through design decisions, data definitions, and proxy variables.

The concept of algorithmic bias also entered the mainstream—alongside the popularization of AI products—as investigations of actual deployed algorithms garnered public attention. In 2016, a team of investigative journalists from ProPublica unveiled how an algorithm that was widely used across the United States for

recidivism was biased against racial minorities (Angwin et al., 2016). Work by Safiya Umoja Noble (2018) made breakthroughs establishing that search engines, far from being a neutral tool, often reproduce systemic racism, sexism, and economic inequity. Ruha Benjamin (2019) further argued that algorithmic bias towards racial minorities is not an accident and a byproduct of design but often an inbuilt feature masked by the language of progress, efficiency, or innovation. Empirical investigations of real-world deployed systems had significant impact in the term "bias" becoming a well-known phenomenon, particularly outside academic research. Buolamwini and Gebru (2018) evaluated three commercial facial recognition technology (FRT) products, demonstrating disparate performance based on gender and skin tone. In popular media, scandals such as the Google Photo App that labeled Black people as "gorillas" (Dougherty, 2015), Amazon's recruitment algorithmic that penalized women (Dastin, 2018), and the wrongful arrest of Robert Williams, a Black man misclassified by FRT (Mullen & Herrera, 2024), all contributed towards the popularization of bias in public understanding. As AI systems increase in scale, sophistication, and modality—especially with the emergence of generative AI—the ways and nuances in which they encode bias are still underexplored [see Large Language Models].

Core concepts

Although bias can arise at any and all stages of the AI pipeline and lifecycle, the most commonly studied types of bias focus on *bias from data and its processing* (i.e., bias that might occur in both the content of the data and its processing pipeline) and *bias from model training and evaluation* (i.e., bias that might be introduced by training architecture and dynamics and optimization choices or evaluation methodology). *Fairness in machine learning* is a core concept and a broad umbrella that encompasses measurement metrics and various technical tools and approaches to address bias.

Bias from data and its processing

In the data collection and processing phase, bias can arise from underrepresentation, misrepresentation, or hyperrepresentation of certain groups in the training data (*representation bias*); from errors such as outdated stereotypes that are often encoded in data, which can go unchecked when labeling, annotating, and categorizing data (*labeling bias*); from inaccuracies and inconsistencies when, for example, features of data are measured or from flawed tools and methods used during data collection (*measurement bias*); from when data is augmented, for example, through synthetic data generation that amplifies stereotypes or distort (from the "ground truth"); and from cleaning and detoxifying data (a process by which data that is considered, for example, denoting toxic language, are removed). *Blocklist filtering* is, for example, one of the methods used to remove content that is presumed to be problematic or toxic. However, like most technical approaches, blocklist filtering comes with drawbacks, for example, resulting in the disproportionate removal of text from and about minority groups, such as removal of language used to communicate about sexual health within LGBTQI+ communities (Dodge et al., 2021).

Bias from model training and evaluation

Bias in model training and evaluation might occur from a number of different processes and choices made. These include the following:

- *Model selection*: Different algorithms process data differently. Using linear models for complex nonlinear relationships, for example, can lead to underperformance on certain groups in the data. Furthermore, some models maybe more prone to overfitting, which could amplify bias.
- *Loss function*: Some loss functions treat all errors equally. However, if the dataset is imbalanced (e.g., underrepresentation of a minority group), the model may prioritize minimizing errors for the majority group, leading to unfair outcomes for underrepresented groups.
- *Over/under fitting*: If a model is trained on a dataset with imbalances (e.g., majority class dominates), it may learn to "memorize" patterns specific to the majority class, which can lead to poor performance on the minority class. Conversely, underfitting occurs when the model fails to learn relevant patterns because it does not have enough data from certain groups. This can happen if the minority class is underrepresented in the training data, and the model fails to capture their specific needs.
- *Model tuning and hyperparameter optimization*: The way hyperparameters are tuned can introduce bias if it leads the model to perform better on one group over another. For example, if hyperparameters are optimized based on overall model performance (accuracy, for instance), the minority class might be overlooked.
- *Model validation and evaluation*: If the validation set is not diverse or balanced, the model's performance may be overstated or inaccurate for certain groups. Using performance metrics such as accuracy, without considerations to all relevant factors, to validate can also lead to biased evaluation (in which one class is much larger than another, for example) giving a false sense of good model performance while in fact failing to account for the distribution of classes in the dataset [see AI Model Evaluation].

Fairness in machine learning

With the core objective of producing equitable outcomes across different individuals and groups, fairness approaches use various techniques to, for example, modify data to "remove" bias before training (for example, reweighting or resampling), modify training algorithms to enforce fairness constraints (for example, adversarial training), or adjust prediction after the model is trained (for example, calibration and thresholding). Fairness metrics—often inspired by or rooted in philosophical theories of distributive justice—aspire to address fair allocation of resources and outcomes. The most common type of fairness in machine learning include the following:

- *Group fairness*, which is premised upon the idea that regardless of the demographic group an individual belongs to, they should receive equal treatment or outcome from a model (Dwork et al., 2011; Hardt et al., 2016). Decision outcomes should thus be independent of any sensitive attributes such as race and gender.

Demographic parity and *equal opportunity* are some of the most commonly used fairness criteria (Dwork et al., 2011).

- *Individual fairness* is based on the idea that similar individuals should receive similar outcomes from a machine learning system (Dwork et al., 2011). Unlike group fairness, which compares outcomes across demographic groups, individual fairness focuses on consistency at the person-to-person level. *Counterfactual fairness* is, for example, a type of individual fairness in which a decision or prediction is considered fair if it would remain the same in the real world and a hypothetical counterfactual world where the individual's sensitive attribute (such as race or gender) had been different but all other relevant factors remained the same (Kusner et al., 2017).

There exist several other criteria, metrics, and alternative approaches. Fairness measures achieve fairness often by making every group worse off or by bringing down better performing groups to the level of the worse off. In rejection of this approach and with the aim of improving outcomes for historically marginalized groups, the concept of *leveling up* (Mittelstadt et al., 2023) has been proposed, whereby systems are designed with minimum acceptable harm thresholds or *minimum rate constraints* (in which a minimum proportion of positive outcomes must be granted to specific demographic groups).

The field has produced countless tools and frameworks geared towards addressing algorithmic bias as well as some of the broader challenges discussed here. These include tools like REVISE, aimed at surfacing bias in visual datasets (Wang et al., 2022); Aequitas, designed to evaluate machine learning models, particularly binary classifiers (Saleiro et al., 2018); and IBM's AI Fairness 360, a large suite with prebuilt datasets and visualization tools (Bellamy et al., 2019). As detailed in the next section, the effectiveness, merit, and utility of these tools is contested.

Notably, *datasheets* and *model cards* remain amongst the most consequential tools and frameworks initiating, popularizing, and subsequently normalizing integration of these tools and practices as industry standard. *Datasheets for datasets* (Gebru et al., 2021) aim to improve transparency through documentation of training data, whereas *model cards* aid the documentation of key information about a model in a consistent and structured manner geared towards increasing transparency and accountability (Mitchell et al., 2019).

Questions, controversies, and new developments

Challenges to implementing fairness

Although several definitions of *fairness* exist within the machine learning literature, there is no clear consensus or universally agreed understanding or definition. *Bias* in natural language processing systems, for example, is vaguely defined, inconsistently applied, and lacks normative reasoning (Blodgett et al., 2020). More importantly, critics have argued that seeking a one-off static solution for structural issues that go under the banner of algorithmic bias via fairness metrics is not only inadequate but also devoid of social or moral

significance and runs the risk of *ethics washing* (false or exaggerated practices or claims while failing to tackle genuine issues; Kasirzadeh, 2022). More fundamentally, the very term *algorithmic bias* is a highly contested term that has been subject to strong *conceptual*, *methodological*, and *epistemological* criticism.

Conceptually, the term *algorithmic bias*, Black feminist and science and technology studies scholars have pointed out, fails to capture wider issues beyond data and algorithms. Issues such as power asymmetries (Brennan et al., 2025) and structural injustices that permeate AI technologies, as well as downstream impacts such as the erosion of fundamental rights including freedom of speech and movement due to pervasive surveillance, are a result of ubiquitous integration of AI into societal infrastructure. Furthermore, unfair and unjust outcomes from deployed algorithmic systems can amount to encoding and amplification of, for example, racism, ableism, misogyny, stagnant stereotypes, or white supremacy with real and immediate real-world impact on actual human beings. Yet, the all-encompassing concept *algorithmic bias* brings in a level of abstractness that makes such issues palatable, removing the immediacy and real-world relevance. This level of abstractness furthermore removes responsibility by making it seem that technological failures or harms from downstream effects are accidental side effects rather than intentional byproducts of oppressive systems (Hampton, 2021).

Seeking technical solutions to complex historical and societal problems, numerous scholars have argued, is not only methodologically limiting but also serves to blur accountability and responsibility. Within the fairness in machine learning approaches, addressing bias primarily consists of mathematical characterization of fairness at the level of data preparation, model learning, or postprocessing, which might result in some incremental change in an algorithm. Broader societal issues that are difficult to quantify and measure—such as desirable social values like human agency and respect for fundamental rights, power and resource asymmetries, and oppressive social structures—that are encoded in and/or intersect with algorithmic systems thus often fall outside the purview of “fairness” due to its narrow methodological scope (Kalluri, 2020). Additionally, within existing fairness paradigms, measuring and operationalizing complex unobservable theoretical constructs such as socioeconomic status via proxies like income remains questionable regarding the reliability and validity of the proxy for directly capturing and measuring the theoretical construct with high fidelity (Jacobs & Wallach, 2021). Similarly, for complex concepts such as fairness and justice, technical interventions render ineffective, inaccurate, and even sometimes dangerously misguided (Selbst et al., 2019).

Epistemic limitations arise from algorithmic formalism, which involves three key orientations: objectivity/neutrality, internalism, and universalism (Green & Viljoen, 2020). Yet, acknowledging, identifying, and addressing societal concerns that arise from algorithmic systems, critics point out, requires pursuing a new mode of thinking outside the bounds of algorithmic formalism, one that is specifically attentive to the internal limits of algorithmic fairness to societal concerns (Green & Viljoen, 2020). The epistemic commitments of neutrality and objectivity of the field of computing leaves no room for analytical scrutiny or deliberation of societal and political questions around data and algorithmic systems. Without the epistemic tools to

systematically investigate underlying values or goals, algorithmic fairness can lead to a narrow range of remedies at the cost of structural reforms and radical interventions.

Perpetual myths

Another controversy pertains the root cause of bias, whereby data is often presented as the sole culprit to the extent that "bias in, bias out" has become a well-known slogan. The idea of "debiasing," one of the proposed remedies for algorithmic bias, has been heavily criticized. One of the strongest empirical arguments against the idea of debiasing comes from natural language processing research demonstrating that debiasing word embeddings provides a superficial solution, giving the illusion of a solution while mostly hiding bias and not removing it (Gonen & Goldberg, 2019).

Despite the relative maturity of the field, the countless ways bias can manifest is not exhaustively studied or known. A robust body of work shows, even when intentional and acknowledging of complexity, interventions at the data or algorithmic level are simply insufficient. For example, even when race is not overtly mentioned nor considered as a variable, large language models (LLMs) can draw from racialized meaning and still make historically racist associations against African Americans based on dialect (Hofmann et al., 2024). What's more, current methods such as "removing" bias via human feedback in training exacerbated racial bias by obscuring racist attitudes instead of mitigating covert racism (Hofmann et al., 2024).

Changing approaches to algorithmic bias

Over the years, new waves of research have emerged, gradually moving away from data- and algorithm-focused solutions and tool development and into a multitude of new directions and emphasis. One trend is a focus on ethics, justice, and accountability from a more holistic analysis of the field. For example, although early fairness work on facial recognition technology focused on measuring bias, accuracy, and performance analysis, a new wave has shifted the conversation towards more fundamental questions about the utility of FRT altogether. This wave is driven by concerns such as the pervasiveness of FRT systems that power surveillance, hyper-surveil the poor and disfranchised, and pose risks to fundamental rights and freedoms regardless of accuracy.

Over the past number of years, audits and evaluations (including of AI models, datasets, and very large online platforms and search engines) have emerged as an effective method and practice for evaluation, assessment, and oversight of AI systems. This stand of development emerged on the backbone of early audits such as *machine bias* (Angwin et al., 2016) and *gender shades (*Buolamwini & Gebru, 2018). Algorithmic audit and evaluation is now a thriving area, with conferences establishing tracks specifically dedicated for audits and evaluation practices over the past few years.

Another domain of research that emerged from the fairness in machine learning via the AI safety community is *AI alignment*. The alignment problem, which has emerged as one of the most popular strands of AI safety

research, is concerned with efforts to align AI systems (often LLMs) to "behave" in line with human intentions and values. This domain of work largely focuses on technical approaches—developing benchmarks, algorithms, and mathematical formalism—to "align" LLMs with human values and preferences. This domain, for the most part, does not deal with normative questions such as the nature of human values or the philosophical, cultural, or epistemological underpinning of "humans," "human behavior," or "values."

Broader connections

For over half a century, fields such as cognitive psychology, behavioral economics, social sciences, philosophy, law, and feminist studies have extensively studied numerous types of human bias such as implicit bias, cognitive bias, and confirmation bias [see Reasoning and Argumentation]. Some of this has resulted in uncovering systematically rooted inequalities, decision-making practices, and inequalities affecting individuals and groups in the real world. These numerous domains and extensive study of bias tend to approach bias from an individualistic manner in which the onus is on the individual (such as through diversity, inclusion, and equity training of individuals) as opposed to structural changes that remove obstacles to marginalized groups. Nonetheless, the past few decades have seen a better awareness of bias, contributing to somewhat more equitable decision-making, including in health care, education, hiring, and public policy [see Eugenic Thinking and the Cognitive Sciences].

The study of algorithmic bias is heavily influenced—directly or implicitly — by cognitive psychology. However, cognitive psychologists have highlighted that these fields' bias evaluation approach and method have faced numerous limitations and, in some cases, suffering crises. Furthermore, inherent differences between human cognition and algorithmic systems means uncritical adoption of human bias evaluation onto algorithms (like implicit association tests) can be misleading, shortsighted, and unhelpful (Crockett et al., 2023).

Acknowledgments

Abeba Birhane is the founder and principal investigator of the AI Accountability Lab (housed in the School of Computer Science and Statistics at Trinity College Dublin), which is supported by grants from the John D. and Catherine T. MacArthur Foundation, the AI Collaborative of the Omidyar Group, Luminate Foundation, European AI & Society Fund, Bestseller Foundation, and The UK Department For Science, Technology and Innovation (AI Security Institute). The author is grateful to Eoin Delaney, Maribeth Rauh, and Harshvardhan Pandit as well as the OECS editorial team for helpful feedback.

Further reading

- Kalluri, P. (2020). Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, *583*(7815), 169. https://doi.org/10.1038/d41586-020-02003-2
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and*

transparency (pp. 59-68). Association for Computing Machinery.
- Winner, L. (1980). Do artifacts have politics? *Daedalus*, *109*(1), 121–136. http://www.jstor.org/stable/20024652

References

- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing ↩
- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671-732. https://doi.org/10.2139/ssrn.2477899 ↩
- Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Natesan Ramamurthy, K., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4:1-4:15. https://doi.org/10.1147/JRD.2019.2942287 ↩
- Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code*. Polity. ↩
- Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. *arXiv*. https://doi.org/10.48550/arXiv.2005.14050 ↩
- Brennan, K., Amba Kak, A., and West, S. M. (2025, June 3). Artificial power: 2025 landscape report. *AI Now*. https://ainowinstitute.org/2025-landscape ↩
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 77-91. ↩
- Crockett, M. J., Bai, X., Kapoor, S., Messeri, L., & Narayanan, A. (2023). The limitations of machine learning models for predicting scientific replicability. *Proceedings of the National Academy of Sciences*, *120*(33), e2307596120. https://doi.org/10.1073/pnas.2307596120

↩

- Dastin, J. (2018, October 10). Insight - Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

↩

- Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., & Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv*. https://doi.org/10.48550/arXiv.2104.08758

↩

- Dougherty, C. (2015, July 1). Google Photos mistakenly labels black people “gorillas.” *New York Times*. https://archive.nytimes.com/bits.blogs.nytimes.com/2015/07/01/google-photos-mistakenly-labels-black-people-gorillas/

↩

- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness through awareness. *arXiv*. https://doi.org/10.48550/arXiv.1104.3913

↩

- Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330-347. https://doi.org/10.1145/230538.230561

↩

- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92. https://doi.org/10.1145/3458723

↩

- Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 609–614). Association for Computational Linguistics.

↩

- Green, B., & Viljoen, S. (2020). Algorithmic realism: Expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 19-31). Association for Computing Machinery.

↩

-

↩

- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *arXiv.* https://doi.org/10.48550/arXiv.1610.02413

↩

- Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature*, *633*(8028), 147-154. https://doi.org/10.1038/s41586-024-07856-5

↩

- Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. *arXiv*. https://doi.org/10.48550/arXiv.1912.05511

↩

- Jain, R. K., Chiu, D. M. W., & Hawe, W. R. (1984). A quantitative measure of fairness and discrimination. *arXiv.* https://doi.org/10.48550/arXiv.cs/9809099

↩

- Kalluri, P. (2020). Don’t ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, *583*(7815), 169. https://doi.org/10.1038/d41586-020-02003-2

↩

- Kasirzadeh, A. (2022). Algorithmic fairness and structural injustice: Insights from feminist political philosophy. *arXiv*. https://doi.org/10.48550/arXiv.2206.00945

↩

- Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *arXiv*. https://doi.org/10.48550/arXiv.1703.06856

↩

- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *arXiv*. https://doi.org/10.48550/arXiv.1810.03993

↩

- Mittelstadt, B., Wachter, S., & Russell, C. (2023). The unfairness of fair machine learning: Leveling down and strict egalitarianism by default. *Michigan Technology Law Review*, 1-49. https://ssrn.com/abstract=4331652

↩

- Mullen, A., & Herrera, S. (2024, June 28). Civil rights advocates achieve the nation’s strongest police department policy on facial recognition technology. *ACLU Michigan*. https://www.aclumich.org/press-releases/civil-rights-advocates-achieve-nations-strongest-police-department-policy-facial/

↩

- Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

↩

- Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. *arXiv.* https://doi.org/10.48550/arXiv.1811.05577

↩

- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019, January). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 59-68). Association for Computing Machinery.

↩

- Wang, A., Liu, A., Zhang, R., Kleiman, A., Kim, L., Zhao, D., Shirai, I., Narayanan, A., & Russakovsky, O. (2022). REVISE: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, *130*(7), 1790-1810. https://doi.org/10.1007/s11263-022-01625-5

↩

- Winner, L. (1978). *Autonomous technology: Technics-out-of-control as a theme in political thought*. MIT Press.

↩

- Winner, L. (1980). Do artifacts have politics? *Daedalus*, *109*(1), 121–136. http://www.jstor.org/stable/20024652

↩

- Winograd, T., & Flores, F. (1986). *Understanding computers and cognition: A new foundation for design*. Ablex.

↩