Uncoupling Inequality:
Reflections on the Ethics of Benchmarks for Digital Media

Anne L. Washington
New York University
washingtona@acm.org

Lauren A. Rhue
Robert H. Smith School of Business
University of Maryland
lrhue@umd.edu

Lisa Nakamura
American Culture Department
University of Michigan
lnakamur@umich.edu

Robin Stevens
Annenberg School for Communication & Journalism
University of Southern California
robinste@usc.edu

Abstract

Our collaboration seeks to demonstrate shared interrogation by exploring the ethics of machine learning benchmarks from a socio-technical management perspective with insight from public health and ethnic studies. Benchmarks, such as ImageNet, are annotated open data sets for training algorithms. The COVID-19 pandemic reinforced the practical need for ethical information infrastructures to analyze digital and social media, especially related to medicine and race. Social media analysis that obscures Black teen mental health and ignores anti-Asian hate fails as information infrastructure. Despite inadequately handling non-dominant voices, machine learning benchmarks are the basis for analysis in operational systems. Turning to the management literature, we interrogate cross-cutting problems of benchmarks through the lens of coupling, or mutual interdependence between people, technologies, and environments. Uncoupling inequality from machine learning benchmarks may require conceptualizing the social dependencies that build structural barriers to inclusion.

1. Introduction

Are large-scale machine learning benchmarks ethical? Benchmarks accelerated computer science research by scientifically tracking performance improvements in algorithms. ImageNET, for example, is an annotated open data set designed to advance machine learning models [1]. Critical scholars continually raise the alarm that analysis of social and digital media using these systems perpetuates inequality and reproduces disadvantages for some populations over others [2, 3, 4].

The ethics of deploying questionable machine learning benchmarks in active systems is particularly vivid as the COVID-19 pandemic illuminated the need for information infrastructures to expertly handle medical [5, 6] and patterns in racial digital media [7]. Despite ample evidence of bias in machine learning benchmarks [8, 9, 10, 11], the chronic reasons for these failures remain unclear.

This paper explores the questionable ethics inherent in machine learning benchmarks from a socio-technical management perspective with insight from public health and ethnic studies in a joint project that seeks to model inquiry across disciplines.

The first authors, trained in management information science, reached out to social science and humanities scholars for response and reflection on the initial concept. In the spirit of shared interrogation, we seamlessly represent those conversations in this collaborative article between computer scientists, social scientists, and scholars of Black and Asian American studies. Our collective inquiry expands a conversation on ethics across field boundaries.

To interrogate machine learning benchmarks as structural barriers, we first situate them as failed information infrastructure and introduce the concept of coupling.

Coupling, or a theory of interdependence between elements, asserts that human, systems, environments, and technologies are linked together in mutual relationship to each other. Next, we trace the impacts of benchmarks in the context of mental health assessments of Black teens. Public health and medicine

URI: https://hdl.handle.net/10125/79686
978-0-9981331-5-7

are particularly vulnerable to failures of machine learning methodologies. Then a humanistic reflection situates coupling dynamics within theories of race with examples of anti-Asian hate on social media. We ask whether ethical problems in benchmarks, taken collectively, are indicators of an unhealthy ecosystem of dependencies. Finally, we connect these observations to the literature on catastrophic systems failure. Organizational sociologists explained physical infrastructure failures by examining the complex interdependence between systems or tight-coupling [12, 13]. We question whether the continued reliance on large-scale scientific benchmarks to evaluate digital and social media poses an equal possibility of catastrophic errors.

We conclude with observations on the ethics of analyzing digital and social media with flawed benchmarks.

2. Information infrastructure of open data benchmarks

Scientific progress in data science is based on evaluating new algorithmic models against the same open data set. Benchmarks are open data sets used within research communities to measure progress in model improvement. An early benchmark for natural language process benchmark, the Brown Corpus [14] was a carefully curated set of texts that took years to annotate. Supervised machine learning continues in this tradition by relying on labeled data sets that establish the "gold standard" to identify observable patterns through statistical calculations. Machine learning heavily relies on annotated benchmarks to build some learning classifiers [15].

Commercial systems may rely on open benchmarks to track performance metrics moving experimental benchmarks into operational systems. Although these datasets may be valuable for benchmarking predictive models, the practice amplifies oversights that are likely to proliferate from test systems to high impact systems [16]. Organizations use these benchmarks as high quality goals [17] instead of experimental baselines, leading to systems that are overly optimized for the benchmarks' narrow universe. Judging by the number of published articles engaging in this practice [18], it seems these datasets are also encouraged by editors. The reuse of these benchmarks is part of the wider trend in the reuse of scientific research data [19].

Publishers of scientific journals seem to encourage the reuse of these data sets, judging by the number of published articles engaging in this practice [18]. Any problem within these Internet platforms is exacerbated when it is integrated into a benchmark that serves as critical information infrastructures in both research and commerce [20, 21].

Benchmarks serve as information infrastructure because they organize what we know and what we can learn. This poses a problem because the data collection process for machine learning benchmarks perpetuates, reinforces, and scales socio-historical patterns of exclusion or negative associations [22]. These problems are not news to benchmark designers and may be simple listed as limitations. For example, one data paper extensively described the undesirable associations between Muslims and Islam [23]. After a detailed critique of ImageNet [9, 24], ImageNet published a similar article outlining similar limitations and solutions [25].

Data size is assumed to measure representativeness, which is not uniformly true when dealing with historically under-represented groups. The emphasis on size was clear in the wide-spread use of "big data" between 2012-2017. A process that prioritizes data size assumes that the population prevalence in the digital space reflects the populations' prevalence offline. The reliance on data size privileges the majority over the minority in the training and testing of models through a quantitative approach. Furthermore, larger size in a historical dataset is an indicator of older rather than newer concepts, yielding overemphasis of older concepts and marginalization of some populations.

Many benchmarks are generated from freely available data on public Internet sites. Yet, online web data prioritizes people with Internet access, who are typically more affluent and educated. For instance, the composition of historical figures on Wikipedia tends to be overly male [26], which could reflect the population of Wikipedia contributors. A large-scale language model, General Pre-trained Transformer 3, GPT-3, was built by gathering words on Reddit and Wikipedia as representative language [8, 23].

Benchmarks serve as critical information infrastructure not only for experimental comparisons but also as the standard for commercial viability, and scientific publishing.

3. Tightly-coupled epistemology

Machine learning benchmarks prefer large sources, rely on data scrapped from public Internet sites, and become de-facto standards across contexts. We question the ability to make meaning in the presence of tight dependency between these characteristics. Large data sets privilege majority voices and public Internet sites are often dominated by male voices [27]. Benchmarks for commercial contexts must be more resilient than research experiments for scientific publishing [28]

Epistemology, or ways of knowing, using these data sources will inevitably create a circular logic about who is and is not present online. Management science theorists refer to this as coupling.

Coupling describes how dependent and responsive people, systems, and technologies are to each other. It is a mechanism of linkage in socio-technical systems. The more systems are interconnected or "tightly coupled", the more likely failure will reproduce throughout the system. Tightly coupled items immediately react to each other because the items are fully interdependent, to the point that there is no room for slack in operations [29, 12].

Tight coupling amplifies small mistakes or errors in assumptions [30]. It also shows how a single problem can magnify exponentially. Although tightly coupled systems are efficient under optimal and ideal conditions, they can spark disaster in unexpected situations. Given that most of these systems assume a majority member of the population [31], deploying them for use with non-majority populations points towards inevitable error.

Benchmarks are not only coupled to one side of reality but they are also coupled to each other. The reuse of digital material across contexts makes projects built on these infrastructures tightly coupled to each other. One popular image dataset, ImageNET [1], is a visual database of tagged images scrapped from Internet websites as representative photos tagged using the WordNET [32] corpus. ImageNET is therefore highly reactive to decisions made in WordNET.

This framing explains the questionable ethics of benchmarks that analyze digital and social media. We believe that to uncouple inequality from these systems, it is necessary to highlight the networks of dependency within and between benchmarks. As academic exercises, these benchmarks benefit research through standardization but reliance on datasets built on free Internet sites has limitations and likely invisible coupling of assumptions about representativeness.

The following two sections contain discussions about the application and practice of the tight-coupling of digital and social media benchmarks in the analysis of historically under-represented groups in the United States. Together, these two insight commentaries provide social science and humanities perspectives on coupling and its implications for ethics of machine learning benchmarks.

4. Public health benchmark dependencies

Public health and medicine are particularly vulnerable to the unanticipated consequences of computational benchmarks. Because most public health researchers lack the tools to fully assess the quality, limits, and threats of machine learning (ML) predictive algorithms, these methods are accepted as nearly totally valid. Many social scientists assume that computation methods have no flaws, no interdependence, and have excellent, representative training data.

In practice, researchers in these fields are not prepared to detect when bias or errors occur. As a result, replicated bias does not become evident until disparities occur in practice, if at all [33]. Once bias due to incorrectly-coupled word and image associations become incorporated into the system, they become reified in practice and thus very difficult to remove.

Benchmarks that are tightly coupled, as the first authors illuminate, pose a central threat. The seminal assumption is that the knowledge discoverable in the sampled data reflects the larger population and thus the patterns discovered from the data would generalize to new data and potentially to a new context. It is this foundational assumption that underpins this critique of algorithms and machine learning in public health practice.

In every research methods class, students learn to interrogate the sample – Who is included? How did we identify them? Who was excluded? Who did we fail to reach? We ask the questions to help identify threats to our study's validity, and to assess how closely sample data approximates the true population. The scale of big data – millions of users, billions of data points--is often accepted as a true approximation of the population. After all, how could a million twitter users not represent the population? The size of the data becomes the smoke and the method of collection, a machine, is the mirror, which in tandem hides real sampling biases. These are the same sources of sample bias that are present in all social science research and warrant interrogation and an accounting of resulting limitations.

When communities are rendered invisible in the computation methods, whether due to crudeness of the tools (e.g. race/ethnicity prediction algorithms) or the lack of integration of community members in the research as advisors or experts, the erasure is easily amplified. Because segments of the population such as Black teens are not identified with the correct keyword or last name identifier they are rendered invisible. We further marginalize the voices and experiences of these parts of the community when their keywords are not included as part of an initial training data set.

For example, language models built to identify depressive symptomology use keywords like “sad” and “upset.” These keywords are typically drawn from samples of white cis and heterosexual populations. However, in our community based work with Black teens, we find the use of the words “sick” and “tired” to be associated with depressive symptomology. If a depression classifier excludes the keywords of Black teens, mental health distress is greatly underestimated in this population. The bias is easily replicated and difficult to identify because without foreknowledge of diverse populations and clarity of how the language model was derived, practitioners are left “not knowing what they don’t know.”

Computational methods are research tools that can and must be scrutinized for threats to validity and limitations. However, without expertise in computer science (CS) and computational methods, public health researchers look to other markers for evidence of methodological accuracy. These markers of a “gold standard” may be signified by 1) a publication record, 2) the successful marketing and sale of a product or commodity, 3) the adoption of these tightly coupled yet racially exclusionary word and image sets by other researchers are less vulnerable to critique because we still lack effective tools to accurately measure harms, 4) or in the promise of improved efficacy in finding the population, making better predictions, or being more broadly representative.

Does it make sense to look to the CS community to offer guidelines, standards or ways to help non-experts identify the strengths and weaknesses of computational approaches before we integrate them into our praxis? Our interdisciplinary perspectives in this paper models an approach to enrich data and data practices by making them more sensitive to lived experience, more accurate, and less or differently tightly-coupled.

The pandemic underlined the immediacy of public health. We understand that it may not make sense in moments of crisis to ask that all data sets be discarded when they reduplicate tightly-coupled but inaccurate word and image links, given that health services must be provided and decisions to care for actual people must be made. Given that our work is pragmatic, on the ground, and embodied, public health researchers have much unique data to offer that can articulate to already-streamlined processes. Adding the keywords to signify depressive symptomology like “sick” and “tired” along with “sad” and “upset” diversifies coupling processes without doing away with them altogether.

Ultimately, practitioners in these fields view computational methods such as artificial intelligence, machine learning, and natural language processing as the proverbial black box, and this is dangerous. This box not only hides and amplifies bias, it obscures methodological limitations, the topic of this paper’s first authors. The label on the Box touting “Big Data” promises greater levels of objectivity, generalizability and efficacy than other social science research methods. These promises are rarely delivered.

5. Decoupled data, noise and Asian-American racial bias in social media

In this rich provocation, the first authors ask whether machine learning benchmarks are ethical? They note how machine learning benchmarks amplify existing biases in WordNet and ImageNet such as the association between the word “terrorism” and “Islam” [23].

Tightly coupled benchmarks unevenly distribute risk on a massive scale and have done so since the adoption of predictive models for criminal justice and health care, facial recognition systems, and the use of big data to make decisions about people’s lives. As the first authors point out, this is inevitable because the systems are designed to be “efficient,” that is to say, use commonly-available models and data in common, to depend upon or couple tightly with each other. Though gender and racial biases are built into the words and images that feed these systems they are not vetted, removed, or revised beforehand; they reflect the concerns and lack of concern of their original creators who as the authors note could not have known how much these associations (between “dark skin” and “man/woman,” or “high healthcare expenditure” and “sick”) would empower men and white people and disempower women and people of color.

The discussion on public health asks us to imagine how people of color can generate new data that is valuable precisely because it is at odds with tightly coupled systems and models. This is less a plea for diversity for its own sake, in the service of ethical behavior and a just society, but rather an argument grounded in what we already know about systemic failures, that is to say, that they are inevitable given a tightly coupled system that uses common models.

Admitting that algorithms discriminate against the same people that health care, educational, financial, and other systems have always discriminated against mean admitting that racism is systemic, acceptable, and constitutive.

Assuming that algorithmic decision making systems are broken out of the box requires that designers and users demand the opposite of what they are given and told they need: systems that are more complex rather than less, data that comes from diverse and at times incongruent sources, and non-standard processes. All of this is counterintuitive during a pandemic. It is precisely during times of crises that racialized failures become rapidly normalized, however.

Both of Lisa Nakamura's parents grew up in internment camps established during World War II to separate and punish Japanese Americans for their racial and ethnic identities. Her mother's family spent three years in the Granada War Relocation Camp in Amache, Colorado, where 7,000 people were kept behind barbed wire between 1942 and 1945. Her father's family was sent to Heart Mountain Camp near Cody, Wyoming. Imprisoning U.S. citizens because of their racial identity became normal because as Wendy Chun writes, "networks presume and prescribe homophily—that birds of a feather flock together, that similarity breeds connection. A banal and therefore dangerous notion of friend becomes a synonym for neighbor: segregation becomes naturalized and hatred becomes love [9]. How do you show you love the same? By fleeing when others show up."

Racism is the opposite of unexpected system failures, as the literature on racial capitalism shows, it is an integral part of our past and present economy, backstopped and founded upon slave, coolie, bracero, and other unfree waged, un-waged, and reproductive labor [20].

How is it the case that the word "chink" and "flu" together do not trigger filtering systems when used on Twitter? Even though we have reported multiple incidents of these specific words used together to target and stigmatize Asians and Asian Americans in the COVID-19 pandemic, they recur on Twitter frequently, and even more frequently after public efforts to curb anti-Asian violence appear in the news or in governmental projects. These two words have been coupled together by xenophobic and racist cultural practice that our contemporary data science regime has not yet evolved to manage.

What might an un-coupling or anti-coupling project look like from the perspective of computer science, ethnic studies, or rhetorical studies? Asian American Studies scholars in particular need this research to understand how anti-Asian hate speech flourishes on social media networks during moments of crisis.

During the same historical moment that many of the unfair algorithms the authors discuss were implemented and scaled up, pundits and scholars celebrated the Internet as an engine that produced "cognitive surplus," and a direct path to "organizing without organizations" [34]. This research invites us to consider whose ends are served when organizations metastasize and reproduce bias during the very moment that they are declared "over." We look forward to the first authors' integrating these concepts on the pandemic's disproportionate negative effects due to tightly coupled and too-hastily (yet too slowly) implemented decision systems.

6. Infrastructure failures

Past industrial accidents inform our understanding of machine learning benchmarks as failures. In the next section, we consider how coupling explained large-scale industrial failures and how this connects to the ethics of machine learning benchmarks.

6.1. Tight coupling in industrial failures

A number of catastrophic industrial incidents in the late twentieth century led to the development of theories about infrastructure failures [35]. In 1979 the Three Mile Island Nuclear Power plant in Pennsylvania suffered a catastrophic environmental disaster [29]. In 1982 a space mission exploded killing all astronauts onboard [13]. Attention to cascading effects lead organizational sociologists to develop the concept of coupling. Problems might remain unnoticed and unresolved until a series of interconnected incidents revealed the hidden dependencies. Importantly, these theoretically rich stories attuned to what ignited the incident and how the catalysts were viewed prior the incident.

Most of industrial failures could not be attributed to a single decision. Petroski [36] makes the distinction between two types of decisions with a negative impact: an error or a mistake. The failure was not caused by a single mistake, i.e., a decision that deviated from a known standard. A single mistake can be addressed by requiring compliance to a standard. The failure was not caused by a single error, i.e., a choice based on a wrong assumption. A single error can be avoided by correcting faulty assumptions. Multiple accumulative problems were fundamental to the definition of an industrial failure.

One problem may be trivial on its own but as part of a wider complex system it may create an unanticipated interaction. The cumulative nature of systems intensifies the difficulties in avoiding failure.

6.2. Pandemic failures in health and race

The pandemic highlighted many failed dependencies without society. The above chronicles of research in public health and ethnic studies paid particular attention to how the normative construction of benchmarks influenced the sub-standard outcomes for others in the current moment.

Public Health. The first section described the applications of computational tools by public health researchers. The global pandemic has accelerated the adoption of computational methods to understand public health, further integrating statistical models and crisis-motivated decision making. Yet in the rush to understand the impact of the pandemic, public health researchers may overlook assumptions about the representativeness of the data. Scraping social media or building an Internet-derived dataset may lock in assumptions that are inappropriate across all public health settings and outcomes.

Ethnic Studies. The second section described the historical context of racial inequality, highlighting additional vulnerabilities as machine learning grapples with the changing social landscape brought on by the coronavirus pandemic. In tandem with the first section, it argues for more accurate data practices that acknowledge cultural differences emerging from analyzing digital and social media about racialized populations. Specifically, the struggle to mitigate harm against Asian American users of social networks is stymied by the tightly coupled benchmarks that fail to notice the issue. A noisier system might encourage loosely coupled, heterogeneous, data gathering and processing. A loosely coupled process might begin to acknowledge the discrimination inherent in data science's primal methods and sources that privilege efficiency.

6.3. Benchmarks as infrastructure failures

Systems failure theory sheds light on assumptions of interdependence and tight coupling that could help to explain the distribution of risk across population groups.

The commentaries related to public health and social media highlight the commonality of this problem across contexts. Despite the significant differences in context, similar populations can be marginalized due to the technology-mediated and convenience-focused data collection. If researchers do not exert a concerted effort to capture representative data, and discuss their insights with a diverse population, then they may never realize the flaws in their data. Without an effort to discuss data collection with researchers who specialize in public health for marginalized communities, researchers gathering social media to predict depression may never realize that Black teenagers use "sick" and "tired" as compared to white teenagers using "sad" and "upset". Thus, the tight-coupling of data sources underscores the importance of an interdisciplinary approach rather than a computer science led effort, especially as it relates to language [37, 8].

The problems of machine learning benchmarks are not unique one-time errors but instead reproduce structural inequality. Understanding machine learning benchmarks as failed information infrastructure is an opportunity highlights the threat of large-scale infrastructure collapse. Current benchmarks designed for scientific publishing of laboratory experiments may not be robust enough for validation across multiple populations.

Any infrastructure designed with the primacy of one population at the expense or apathy towards others invites ethical concerns in both the process and the consequences [38]. The simple mistakes chronicled in over ten years of critical scholarship may be a harbinger for the possibility of a future catastrophic failure. A systems failure perspective shifts the conversation from specific instances of harm to recognition that collectively these are risk indicators for the whole system. In particular, the system-wide failure may trigger unevenly distributed consequences for those populations unrepresented by the benchmarks and for whom the system does not see.

The coupling literature reveals the interplay between these elements and provides a path towards illustrating current situations and questioning paths towards improvement. It invites us to consider how negative outcomes can both occur and scale within entanglements of benchmarks, people, and systems.

7. Discussion and provocations

As we build a society reliant on only a few flawed large-scale machine learning benchmarks, the small errors currently seen in population sub-groups may scale exponentially to pose significant risks.

As data-driven models pervade more areas of society, machine learning benchmarks move from an esoteric issue to a systemic one. The narrow qualifications for data scientists, focusing on computational prowess without ethical training or humanities-based knowledge [11], underlie much of the current conundrums associated with prioritization of algorithmic optimization using tightly-coupled benchmark datasets instead of prioritization of risk reduction and inclusivity.

By incorporating an interdisciplinary approach, data scientists can examine the consequences and choices through a multidimensional lens--statistics and society--rather than a narrow focus on expediency and computational methods.

Because populations access digital and social media at different rates, digital and social media datasets are often unrepresentative in ways that are predictable to scholars who study society and societal challenges. Plus, these datasets also reflect crowd-based preferences (whether through complaints, responses, or data labels) and thus perpetuate existing societal structures. For instance, the decision not to filter racial slurs related to the pandemic prioritizes the apathy of a majority rather than the dignity and outrage of a marginalized minority. The repetition of similar datasets, e.g., filtering choices across social media platforms, only serves to present an appearance of independence while having a substance of tightly-coupled interdependence.

Time is a major challenge to the adoption of the loosely-coupled and interdisciplinary approach to datasets, particularly in fast-paced fields such as public health. Tightly coupled systems are designed for efficiency and in a crisis, immediate predictions may be necessary to provide care in a crisis. However, the emphasis on immediacy at the expense of inclusion could bifurcate patient care and health outcomes, yielding better decisions for those included in the data and worse decisions for people invisible in the data such as in problems with medical race correction algorithms [33].

Future research could consider how benchmarks help to establish new fields. Understanding how subfields within machine learning differ could be ripe for further investigation as well. Given recent empirical evidence of citations to the DukeMTMC and Labeled Faces in the Wild (LFW) benchmarks [18], the research community is beginning to reflect how these open data sets serve their needs and the public interest.

The second commentary asks a provocative question: rather than loosely coupled datasets, what about un-coupled or anti-coupled datasets? These datasets could capture disparate and complementary information as a means to reduce the unknown unrepresentativeness in benchmarks. Actively seeking smaller or less well-known datasets could reduce the coupling issue and potentially provide more incentives to create a larger repository of smaller datasets instead of assuming that larger datasets are representative. Anti-coupling also raises questions about the focus on computational methods to collect data, particularly Internet-based and digital data.

8. Conclusion

Large-scale open benchmarks facilitate incremental progress on predictions yet they also hold a hidden threat. The popular benchmarks discussed here compound incremental data biases which can lead to increased risk of predictive failure for some populations. Tightly-coupled systems caused catastrophic socio-technical accidents in physical infrastructure and this paper highlighted a similar potential for failure from tightly-coupled datasets. Machine learning benchmarks may have less visible but equally devastating dynamics which can compromise the ethics of researchers and their analysis. As researchers we should diversify our sources and methods of data collection as well as our understanding of how to benchmark and evaluate algorithms. Uncoupling inequality from machine learning benchmarks may require conceptualizing the dependencies that build structural barriers to inclusion.

9. Acknowledgements

Research assistance based, in part, on the first author's funding through National Science Foundation Grant No. #1833119. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

10. References

[1] Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255. Miami, FL: IEEE. doi: 10.1109/CVPR.2009.5206848

[2] Benjamin, R. (2019). Captivating technology: Race, carceral technoscience, and liberatory imagination in everyday life. Place of publication not identified: Duke University Press.

[3] Costanza-Chock, S. (2020). Design justice: Community-led practices to build the worlds we need. MIT Press.

[4] Raji, Inioluwa Deborah, and Joy Buolamwini. 2019. "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products." In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES, New York: Association for Computing Machinery, 429–35. doi: 10.1145/3306618.3314244.

[5] Manchanda, Emily Cleveland, Cheri Couillard, and Karthik Sivashanker. 2020. "Inequity in Crisis Standards of Care." New England Journal of Medicine. doi: 10.1056/NEJMp2011359.

[6] Yancy, C. W. (2020). COVID-19 and African Americans. Journal of the American Medical Association, 323(19), 1891–1892.

[7] Apprich, Clemens, Wendy Hui Kyong Chun, Florian Cramer, and Hito Steyerl. 2019. Pattern Discrimination.

[8] Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In Proceedings of FAAcT the 2021 ACM

Conference on Fairness, Accountability, and Transparency, Virtual Event Canada: ACM, 610–23. doi: 10.1145/3442188.3445922.

[9] Birhane, A., & Prabhu, V. U. (2021). Large Image Datasets: A Pyrrhic Win for Computer Vision? Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 1537–1547.

[10] Buolamwini, Joy, and Timnit Gebru. 2018. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” Proceedings of Machine Learning Research 81: 1–15.

[11] Mitchell, S., Potash, E., Barocas, S., D’Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. Annual Review of Statistics and Its Application, 8(1), 141–163. doi: 10.1146/annurev-statistics-042720-125902

[12] Weick, Karl E. 1976. “Educational Organizations as Loosely Coupled Systems.” Administrative Science Quarterly 21(1): 1–19.

[13] Vaughan, Diane. 1990. “Autonomy, Interdependence, and Social Control: NASA and the Space Shuttle Challenger.” Administrative Science Quarterly 35(2): 225.

[14] Kucera, Henry, and W. Nelson Francis. 1967. Computational Analysis of Present-Day American English (Brown Corpus). Providence RI: Brown University Press.

[15] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. http://arxiv.org/abs/1908.09635

[16] Lambrecht, Anja, and Catherine Tucker. 2019. “Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads.” Management Science 65(7): 2966–81.

[17] Rhue, Lauren. 2019. “Beauty’s in the AI of the Beholder: How AI Anchors Subjective and Objective Predictions.” ICIS 2019 Proceedings. https://aisel.aisnet.org/icis2019/future_of_work/future_work/15.

[18] Peng, K., Mathur, A, Narayanan, A. 2021. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. arXiv:2108.02922 [cs]. (Aug. 2021).

[19] Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

[20] Cotton, Tressie McMillan. (2020). “Where Platform Capitalism and Racial Capitalism Meet: The Sociology of Race and Racism in the Digital Society.” Sociology of Race and Ethnicity. doi: 10.1177/2332649220949473

[21] Stevens, R., Bonett, S., Kenyatta, K., Chittamuru, D., & Bleakley, A. (2019, January 8). Sex, Drugs, and Alcohol in the Digital Neighborhood: A multi-method analysis of online discourse amongst Black and Hispanic Youth. doi: 10.24251/HICSS.2019.261

[22] Bolukbasi, Tolga et al. 2016. “Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings.” In Advances in Neural Information Processing Systems 29, eds. D. D. Lee et al. Curran Associates, 4349–57. http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf .

[23] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language Models are Few-Shot Learners. ArXiv:2005.14165 [Cs].

[24] Prabhu, V. 2021. A study of “A Study of Face Obfuscation in ImageNet.” Medium. https://vinayprabhu.medium.com/a-study-of-a-study-of-face-obfuscation-in-imagenet-d7e7591795a2

[25] Yang, Kaiyu, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. (2021). A Study of Face Obfuscation in ImageNet. https://arxiv.org/abs/2103.06191v2

[26] Reagle, J., & Rhue, L. (2011). Gender Bias in Wikipedia and Britannica. International Journal of Communication, 5(0), 21.

[27] Sugimoto, C. R., Andrejevic, M., Ekbia, H. R., Mattioli, M., Cate, F. H., Burdon, M. (2016). Big Data Is Not a Monolith. Cambridge, MA: MIT Press.

[28] Washington, Anne L. (2016). Interviewing data—The art of interpretation in analytics. Proceedings of the 2016 iConference. doi: 10.9776/16256

[29] Perrow, Charles. 2011. Normal Accidents: Living with High Risk Technologies - Updated Edition. Princeton University Press.

[30] Torralba, Antonio, and Alexei A. Efros. 2011. “Unbiased Look at Dataset Bias.” Proceedings of CVPR 2011 IEEE Conference on Computer Vision and Pattern Recognition, 1521–28.

[31] Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. “Semantics Derived Automatically from Language Corpora Contain Human-like Biases.” Science 356(6334): 183–86.

[32] Miller, George A. 1993. “Wordnet: A Lexical Database for English.” In HLT ’93: Proceedings of the Workshop on Human Language Technology, 409.

[33] Vyas, Darshali A., Leo G. Eisenstein, and David S. Jones. 2020. “Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms.” New England Journal of Medicine 383(9): 874–82.

[34] Shirky, Clay. (2009). Here Comes Everybody: the Power of Organizing Without Organizations. Penguin.

[35] LaPorte, Todd R., and Paula M. Consolini. 1991. “Working in Practice But Not in Theory: Theoretical Challenges of `High-Reliability Organizations’.” Journal of Public Administration Research and Theory 1(1): 19–48.

[36] Petroski, H. 1992. To engineer is human : the role of failure in successful design. Vintage Books.

[37] Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of “Bias” in NLP. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. http://arxiv.org/abs/2005.14050

[38] Mittelstadt, B. D., & Floridi, L. (2016). The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. Science and Engineering Ethics, 22(2), 303–341. doi: 10.1007/s11948-015-9652-2

[39] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163. doi:10.1089/big.2016.0047

[40] Chun, Wendy. (2020) "Net-munity or the Space Between Us...will Open the Future," Critical Inquiry, https://critinq.wordpress.com/2020/05/20/net-munity-or-the-space-between-us-will-open-the-future/

[41] D'Amour, Alexander et al. 2020. "Underspecification Presents Challenges for Credibility in Modern Machine Learning." http://arxiv.org/abs/2011.03395.

[42] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. 1996. "The KDD Process for Extracting Useful Knowledge from Volumes of Data." Communications of the ACM 39(11): 27–34.

[43] Heaven, Will Douglas. 2020. "The Way We Train AI Is Fundamentally Flawed." MIT Technology Review. https://www.technologyreview.com/2020/11/18/1012234/training-machine-learning-broken-real-world-heath-nlp-computer-vision/ (April 26, 2021).

[44] Kleinberg, J., Mullainathan, Sendhil, & Raghavan, Manish (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. Proceedings of Innovations in Theoretical Computer Science (ITCS 2017). https://arxiv.org/abs/1609.05807

[45] Learned-Miller, E., Huang, G. B., RoyChowdhury, A., Li, H., & Hua, G. (2016). Labeled Faces in the Wild: A Survey. In M. Kawulok, M. E. Celebi, & B. Smolka (Eds.), Advances in Face Detection and Facial Image Analysis (pp. 189–248). Cham: Springer International Publishing. doi: 10.1007/978-3-319-25958-1_8

[46] Nakamura, L., & Chow-White, P. (2013). Race After the Internet. Routledge.

[47] Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." Science 366(6464): 447–53.

[48] Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. "Auditing Algorithms." ICA International Communication Association 2014 Data and Discrimination Preconference: 1–23.

[49] Shrestha, Yash Raj, Shiko M. Ben-Menahem, and Georg von Krogh. 2019. "Organizational Decision-Making Structures in the Age of Artificial Intelligence." California Management Review 61(4): 66–83.