# Bigger Isn't Better: The Ethical and Scientific Vices of Extra-Large Datasets in Language Models

TRYSTAN S. GOETZE and DARREN ABRAMSON, Dalhousie University, Canada

The use of language models in Web applications and other areas of computing and business have grown significantly over the last five years. One reason for this growth is the improvement in performance of language models on a number of benchmarks — but a side effect of these advances has been the adoption of a "bigger is always better" paradigm when it comes to the size of training, testing, and challenge datasets. Drawing on previous criticisms of this paradigm as applied to large training datasets crawled from pre-existing text on the Web, we extend the critique to challenge datasets custom-created by crowdworkers. We present several sets of criticisms, where ethical and scientific issues in language model research reinforce each other: labour injustices in crowdwork, dataset quality and inscrutability, inequities in the research community, and centralized corporate control of the technology. We also present a new type of tool for researchers to use in examining large datasets when evaluating them for quality.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; • **Information systems** → *Crowdsourcing*; • **Social and professional topics** → Licensing; Computing profession.

Additional Key Words and Phrases: datasets, language models, computing ethics, epistemology of computing

**ACM Reference Format:**
Trystan S. Goetze and Darren Abramson. 2021. Bigger Isn't Better: The Ethical and Scientific Vices of Extra-Large Datasets in Language Models. In *13th ACM Web Science Conference 2021 (WebSci '21 Companion), June 21–25, 2021, Virtual Event, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3462741.3466809

## 1 INTRODUCTION

One way to describe the history of computing is as a series of pendulum swings between two extremes. On one side, there are the *techno-utopians*, who dream of a post-scarcity society enabled by the widespread adoption of computing technologies that can be customized to every user's needs [33]. On the other side, there are the *techno-capitalists*, who imagine a world where centralized control of computing technologies can wring profit out of every data point that may be collected from end users [34]. We have seen these swings of popular ideology again and again, from the ongoing free/open source vs. proprietary software discourse that began with the earliest personal computers, to the initial hopes for the World Wide Web as a democratizing force that today seem naïve given recurring stories of disdain for privacy and human rights in the pursuit of profit by some of the largest Web-based companies.

In this paper, we are concerned with a recent swing to the techno-capitalist side in the field of language models (LMs). A significant contributor to the success of the modern Web is the rapid rise of natural language understanding models. Since IBM publicly demonstrated the technology's capabilities by showcasing Watson in a *Jeopardy!* exhibition match in 2011, machine learning–driven language processing has become an essential part of Web-based customer service, analytics, healthcare, banking, and other business applications.

However, as a recent critique of LM methodology shows, there are worrying trends in how these models are produced [4]. In the last five years, LMs have been growing dramatically both in terms of the number of parameters and the size of datasets used for training and testing. While larger models have shown significant successes on a number of important benchmarks, the trend towards ever-larger models and datasets comes with significant moral risk. [4] call particular attention to the ethical problems raised by the massive environmental impact and rising financial cost of training large LMs on large datasets, as well as the increased difficulty of determining what data are actually *in* these datasets. We share [4]'s general aims and convictions, and in what follows, we present an expansion of these criticisms of the "bigger is always better" mindset in LM development. Of particular importance to our argument is that ethical and scientific vices come hand-in-hand, particularly given the dependence of large LM development on corporate cloud computing and, in the cases we examine, the microtask economy.

The argument proceeds as follows. In §2, we discuss the trend towards larger datasets in LM development as it relates to LM challenges, paying particular attention to the labour injustices involved in the use of crowdwork. In §3, we suggest that the exploitative working environment of crowdwork combined with the difficulty of scrutinizing large datasets risks creating low-quality datasets whose flaws go unnoticed. In §4, we draw out epistemological worries with the "bigger is always better" paradigm, arguing that the increased financial costs of large datasets, and the accompanying increase in corporate power in this research area, will be damaging to the LM research community and to the results it produces. Along the way, we make several suggestions for mitigating ethical concerns; in §5, as a partial way of mitigating epistemological problems, we introduce a tool we call `nlp-data-explorers` for researchers to examine large datasets when evaluating their quality. §6 concludes.

## 2 THE HUMAN COST OF "HUMAN INTELLIGENCE"

A limitation of [4] is that they confine their critique to LM training datasets that have been crawled from existing text corpora on the Web. But the trend towards larger datasets has also influenced the development of LM *challenges*. Because these challenges are typically narrowly defined tasks that are easy for humans but difficult for LMs, they cannot be created simply by assembling a massive collection of publicly accessible textual data. Instead, researchers define a formula for test questions, then either create a set of problems themselves, or assign the task of creating problems to microtask workers through services such as Amazon Mechanical Turk.

Consider CommonsenseQA, a challenge designed to test an LM's "commonsense" understanding [35]. (It is beyond the scope of this paper to discuss, but it is worth noting that CommonsenseQA does not engage with [7]'s well-known argument that computer systems can never achieve commonsense.) To develop the CommonsenseQA challenge, the researchers engaged crowdworkers to create over 12,000 multiple choice questions based on the links between concepts in ConceptNet [32]. For example, from the link between the concepts *river* and *waterfall*, a crowdworker might create a question like, "You would expect to find a waterfall at the end of a what?", with "river" being the correct answer. While we concentrate on the case of CommonsenseQA, some LM challenges have developed even larger datasets: e.g., the WinoGrande challenge [28], which tests an LM's ability to handle ambiguous referents in Winograd schemas (see [19]), relies on a dataset of about 44,000 crowdworker-generated problems.

This approach is problematic, as the use of crowdwork comes with well-documented moral risk [11, 13, 16, 17, 21–23, 29, 30]. Crowdworkers are generally extremely poorly paid for their time; ineligible for benefits, overtime pay, and legal or union protections; vulnerable to exploitation by work requesters; likely to lose wages to "downtime" spent looking for decently paying work; and subject to deceit, obfuscation, and intimidation from the platforms that mediate between them and work requesters. Moreover, many crowdworkers end up trapped in this situation due to a lack of jobs in their geographic area for people with their qualifications, compounded with other effects of poverty.

Some researchers have suggested potential remedies to this moral risk. For example, building on calls [31] to pay crowdworkers at least minimum wage, [38] suggest one relatively simple intervention that they call "Fair Work." Their approach enables crowdworkers to report their actual time spent on microtasks, allowing their wages to be topped up to a "fair" rate of US $15/hour by the researcher. It is unclear how widely such principles have been adopted, however; and, as we return to below, the increased financial cost may be burdensome for some research groups.

## 3 KNOW YOUR DATA

Even supposing that crowdworkers are fairly paid for their service to computer science, two ethical problems with this research paradigm remain. Firstly, a fair wage is not yet a fair working environment: fairly compensated crowdworkers would still be ineligible for benefits and protections, and subject to intimidation from platform managers. Without sweeping regulatory changes to enforce crowdworkers' labour rights, even researchers who follow best practices are complicit in an exploitative marketplace. Secondly, and more significantly from a scientific standpoint, we suggest that precisely this exploitative arrangement could lead to the production of poor quality datasets, undermining research based upon them.

Concerns about the quality of crowdwork-generated data have been discussed in the social science context, where crowdworker surveys are relied upon for collecting psychological and sociological data. [26, p. 185] found that "workers are diverse but not representative of the populations they are drawn from," with regard to personality, educational background, age, and other demographic markers. This casts doubt on whether challenges such as CommonsenseQA actually capture what can properly be called commonsense understanding. To paraphrase [14], when we build datasets for these challenges, we need to ask, *whose* commonsense and *whose* understanding are we capturing and testing for?

This issue recalls [4]'s worry about the possibility of unreported bias in datasets. A suggestion they make which would apply here is the inclusion of *data statements* [3]. These information slips are presented as appendices to LMs that include information on the linguistic data contained in the dataset, and demographic information on the people who created and annotated the data. A data statement for crowdwork-generated datasets would specify the self-reported demographics of the crowdworkers whose labour produced the data, enabling human researchers or automated tools to scan for the presence of bias.

Data statements only go so far, however, for the working environment of crowdwork is itself in tension with the demands of dataset generation for LM research. In order to be properly composed, the problems that constitute challenges like CommonsenseQA require precise attention to linguistic details. Given the pressures on crowdworkers intrinsic to the crowdwork economy, there is good reason to think that such attention is frequently absent. This shortcoming would be less problematic if the resulting datasets were small enough for researchers to scrutinize for quality before publication, but the desired scale makes such curation impossible. Instead, datasets like CommonsenseQA rely on additional crowdworkers for data validation [35]. However, this solution only re-introduces precisely the same concerns at a higher level.

This underscores another of [4]'s worries, namely, that the contents of large datasets are difficult to examine. While their primary concern is with the presence of bias, overall dataset quality is also difficult to determine when the dataset is sufficiently large. If, as we argue, the nature of the crowdwork economy is in tension with the demands of dataset creation, large challenge datasets like CommonsenseQA could have significant flaws that go unnoticed. If true, these pernicious flaws would undermine claims regarding an LM's performance on the challenge. For example, it would be difficult to tell if a poor score represents a deficiency of the LM or of the dataset.

## 4 EPISTEMOLOGICAL IMPLICATIONS

A further set of problems with large datasets arises from the simple fact that the larger one's LM, and the larger the datasets one feeds it, the more computing power one needs. As [4, p. 9] also argue, a research paradigm with a high financial bar to entry stands to exclude researchers from institutions and countries with limited research funds, further deepening inequality in the research community.

These inequalities represent more than an ethical risk: they also present *epistemological* risks. A research paradigm dependent on financially inaccessible computing resources shuts out citizen scientists whose contributions have historically played pivotal roles in the history of computing. Furthermore, by excluding these outsider contributions and marginalized researchers at less resourced institutions, the "bigger is always better" paradigm can be expected to reduce the diversity of the LM research community, contributing to what [37] call the "diversity crisis" in AI research. And, as philosophers of science have argued for over a century [2, 20, 27], a diverse community of inquiry is necessary to filter out biases that may go unnoticed in a demographically homogeneous group of researchers. We have seen how a lack of diversity in computer science research in particular has led to errors many times before [6, 12, 25]. Machine learning research thus stands to be less objective and, as [36] suggests, LM research in particular stands to be less reliable.

In addition to these mixed ethical-epistemological problems, the financial costs of the "bigger is always better" paradigm increase corporate power in LM research. LM projects are already often dependent on Big Tech firms, such as Microsoft [24], that offer paid cloud computing services to businesses and researchers without the resources or expertise to train and customize machine learning models locally. But regardless of whether the resulting applications are open source, they are not free software [8], because of the centralized control over how the service may be used that Microsoft and other providers maintain. As [18] observe, when corporations retain this kind of power, it impairs the autonomy of users and smaller developers. The "bigger is always better" paradigm thus serves the interests of Big Tech firms as much as it serves the interest of LM research — and perhaps more, since they retain the power to restrict what outsiders can do with their services.

The worry about corporate power in LM research is more than a familiar lament about wealth inequality. This kind of corporate influence has been observed to be damaging to research in other scientific domains. For example, as [5] observes in the context of the pharmaceutical industry, when corporate interests drive research through private research grants, studies that are published tend to favour their donors' interests — e.g. drug efficacy and safety trials are more likely to favour the donor's products — and studies with results opposed to the donor's interests are often suppressed. The recent ouster of two prominent AI ethicists at Google, in part for their contributions to [4], suggests that the same patterns of corporate interference are active in LM research [10].

## 5 EXPLORING DATASETS

The use of large datasets is still probably required for some aspects of LM development and machine learning generally. However, given the concerns we have discussed in this paper, researchers have all the more reason to think carefully about whether large datasets are actually needed to answer their research questions, as [4] also urge. It thus behooves LM researchers to devise methods of mitigating the risks inherent to the creation and use of large datasets.

We have already discussed some strategies for addressing the ethical issues of crowdworker exploitation and data bias. A potential way to address the epistemological risk of quality problems would be to make it easier for researchers


Fig. 1. The initial screen presented to the user in `CQA_Explorer.py`, an `nlp-data-explorer` that taps CommonsenseQA's dataset.

to explore the contents of challenge datasets. To this end, we introduce a type of tool we call `nlp-data-explorers` [1, full code is in the auxilliary files]. Each explorer is an executable python file run from the command line that pulls problems from a dataset, such as CommonsenseQA or WinoGrande, and presents them to the user in a multiple choice test (see Figures 1–3). The user can thereby view a random selection of problems from the dataset, test their performance against the "correct" answers, and compare their scores to an LM by cross-referencing the LM scores reported in publications or leaderboards. With a large enough sample, a coherent snapshot of the dataset as a whole can be captured, and its quality evaluated. Tools like these can supplement data statements and other types of dataset information slips (e.g. nutrition labels [15] or datasheets [9]) by allowing researchers to explore datasets for themselves before using them, or before recommending papers presenting the dataset for publication.

For example, using an `nlp-data-explorer` that taps CommonsenseQA, we were able to find multiple issues that lead us to recommend against using it as a challenge for LMs. Table 1 lists some of the prompts we observed, with the "correct" answer marked in **boldface**. We found items that contain grammatical errors, admit multiple correct

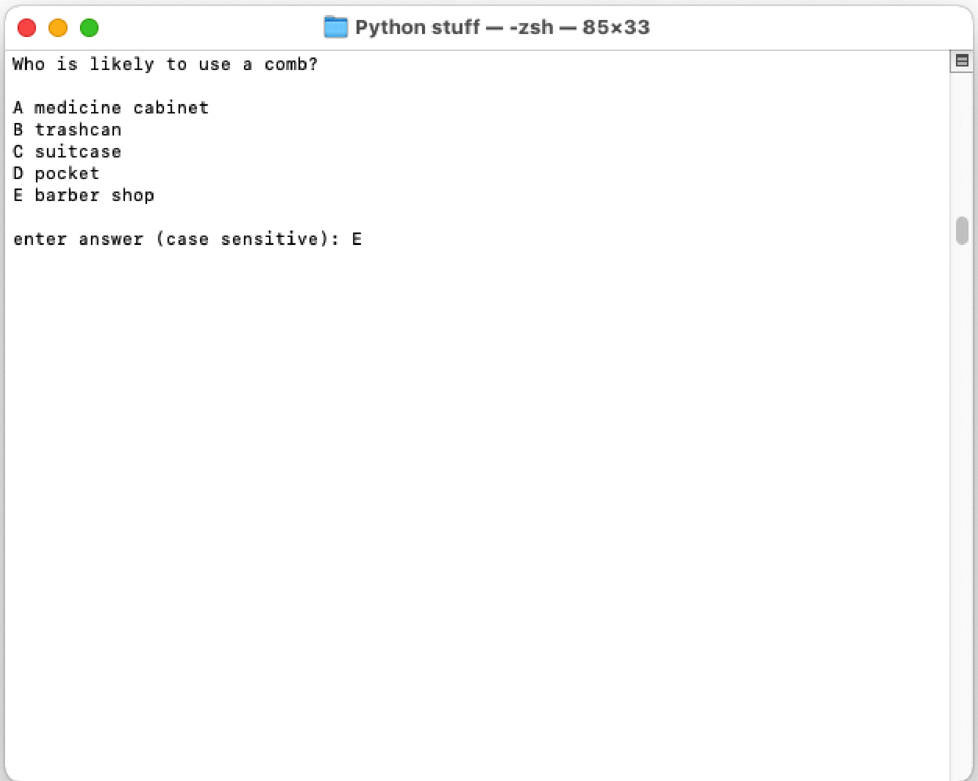


Fig. 2. A question from the CommonsenseQA dataset in `CQA_Explorer.py`.

interpretations, or that have "correct" answers that are inaccurate. These findings corroborate our suspicion that large crowdwork-generated datasets, even those that have been "verified" by additional crowdworkers, may have quality issues. Furthermore, the nature of the errors makes us suspect that CommonsenseQA fails to provide a proper test of commonsense understanding. An LM may fail to answer questions "correctly" because of grammatical errors that lead to mistaken interpretations. Or an LM may fail to determine the "correct" answer because multiple answers are potentially admissible. Or, an LM may choose the "correct" answer merely because it is the only answer whose grammar agrees with the prompt. These issues make it difficult to determine what, if anything, CommonsenseQA measures when testing an LM.

## 6 CONCLUSION

Let's take stock. There are mutually reinforcing ethical and scientific problems with the trend towards ever-larger datasets in LM research and LM applications on the Web and elsewhere. The first set of problems arise from the


Fig. 3. The end screen presented to the user after a session running `CQA_Explorer.py`. After their overall score, for each question the user is presented with (i) the answer key, (ii) the hashcode ID of the question, (iii) the concept from ConceptNet that the question is probing, (iv) the possible answers offered, (v) the question prompt, (vi) the user's answer, (vii) the time in seconds that the user took to enter their answer.

engagement of crowdworkers in the creation of these datasets. Not only is the microtask economy fraught with labour injustices, the working environment so produced raises worries about the quality of datasets created with this method. But given the sheer size of the resulting datasets, they are often difficult for researchers to scrutinize for quality issues. Finally, we contended that a research paradigm desirous of large datasets not only risks pricing out citizen scientists and marginalized researchers, it also actively contributes to the centralization of corporate control in LM research. Such control is not only antithetical to the principles of free software; there is also good reason to think that it will allow large tech firms to push research along directions that suit their business interests over scientific progress or societal interests.

In light of these arguments, we suggest that LM researchers should consider carefully whether creating or processing a large dataset is actually necessary to answer their research questions. We additionally urge research ethics boards

Table 1. Selected problems from CommonsenseQA

| Prompt | Answers | Issues |
|---|---|---|
| She couldn't hide she liked the boy she was talking to, she had a constant what? | (A) Make eye contact<br>**(B) Smile**<br>(C) Another person<br>(D) Listening<br>(E) Compliment | Only solution agrees with sentence<br>Answers not parallel structure<br>Nonsense answers |
| Where might be an odd place to put a washing machine? | (A) Laundromat<br>(B) Wash clothes<br>**(C) Cellar**<br>(D) House<br>(E) Garage | Questionable solution<br>Answers not parallel structure |
| What is a steel cable called a wire rope primarily used for? | (A) Factory<br>(B) Building<br>(C) Winch<br>**(D) Ship**<br>(E) Jumprope | Multiple correct answers |

to become familiar with the ethical and epistemological risks of the use of large datasets in LM research, to require researchers to pay crowdworkers a fair wage, and to require researchers to publish data statements and dataset explorers to accompany their work. These changes will help mitigate the risks we have called attention to, and to nudge the pendulum away from the techno-capitalist extreme.

## ACKNOWLEDGMENTS

The authors' work on this project was supported by a Banting Postdoctoral Fellowship from the Social Sciences and Humanities Research Council of Canada (Goetze), as well as GPU time from Compute Canada and internships funded by the Mitacs Globalink project (Abramson). We also wish to acknowledge that Dalhousie University is located in Mi'kma'ki, the ancestral and unceded territory of the Mi'kmaq. We are all Treaty people.

## REFERENCES


[1] Darren Abramson. 2021. nlp-data-explorer. https://github.com/DarrenAbramson/nlp-data-explorer

[2] Elizabeth Anderson. 1995. Feminist Epistemology: An Interpretation and a Defense. *Hypatia* 10 (1995), 50–84. Issue 3.

[3] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Canada, virtual). ACM, New York, NY, 610–623. https://doi.org/10.1145/3442188.3445922

[5] James Robert Brown. 2002. Funding, Objectivity and the Socialization of Medical Research. *Science and Engineering Ethics* 8 (2002), 295–308. Issue 3.

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research* 81 (2018), 1–15. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

[7] Hubert L. Dreyfus. 1992. *What Computers Still Can't Do.* MIT Press, Cambridge, MA.

[8] Free Software Foundation. [n.d.]. What is free software? https://www.gnu.org/philosophy/free-sw.html

[9] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. 2020. Datasheets for Datasets. arXiv:1803.09010 [cs.DB]

[10] Nico Grant, Dina Bass, and Josh Eidelson. 2021. Google Fires Researcher Meg Mitchell, Escalating AI Saga. *Bloomberg* (2021). https://www.bloomberg.com/news/articles/2021-02-19/google-fires-researcher-meg-mitchell-escalating-ai-saga

[11] M. L. Gray and S Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.* Houghton Mifflin Harcourt, 2019.

[12] Leslie Marie Hampton. 2021. Black Feminist Musings on Algorithmic Oppression. In *Conference on Fairness, Accountability, and Transparency (FAccT '21)*. ACM, New York, NY, 12 pages.

[13] Kotaro Hara, Abi Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey Bigham. 2017. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. arXiv:1712.05796 [cs.CY]

[14] Sandra Harding. 1991. *Whose Science? Whose Knowledge? Thinking from Women's Lives.* Cornell University Press, Ithaca, NY.

[15] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. arXiv:1805.03677 [cs.DB]

[16] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. *The Future of Crowd Work.* Association for Computing Machinery, New York, NY, USA, 1301–1318. https://doi.org/10.1145/2441776.2441923

[17] Tamara Kneese, Alex Rosenblat, and danah boyd. 2014. Understanding Fair Labor Practices in a Networked Age. , 17 pages. https://www.datasociety.net/pubs/fow/FairLabor.pdf

[18] Bradley M. Kuhn and Richard Stallman. 2001/2015. *Free Software, Free Society: Selected Essays of Richard M. Stallman* (3rd ed.). Free Software Foundation, Boston, MA, Chapter "Freedom or Power?", 257–258. https://www.gnu.org/doc/fsfs3-hardcover.pdf

[19] Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd schema challenge. In *KR'12: Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*. Association for the Advancement of Artificial Intelligence, Palo Alto, CA, 552–561.

[20] Helen E. Longino. 2001. *The Fate of Knowledge.* Princeton University Press, Princeton, NJ.

[21] David Martin, Sheelagh Carpendale, Neha Gupta, Tobias Hoßfeld, Babak Naderi, Judith Redi, Ernestasia Siahaan, and Ina Wechsung. 2017. *Evaluation in the Crowd: Crowdsourcing and Human-Centered Experiments.* Springer, Cham, Switzerland, Chapter Understanding the Crowd: Ethical and Practical Matters in the Academic Use of Crowdsourcing, 27–69. https://doi.org/10.1007/978-3-319-66435-4_3

[22] David Martin, Benjamin V. Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a Turker. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 224–235. https://doi.org/10.1145/2531602.2531663

[23] B. McInnis, D. Cosley, C. Nam, and D. Leshed. 2016. Taking a HIT: Designing around rejection, mistrust, risk, and workers' experiences in Amazon Mechanical Turk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* ACM, San Jose, CA, 2271–2282.

[24] Microsoft. [n.d.]. Microsoft Artificial Intelligence. https://microsoft.com/ai

[25] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York University Press, New York, NY.

[26] Gabriele Paolacci and Jesse Chandler. 2014. Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science* 23, 3 (2014), 184–188. https://doi.org/10.1177/0963721414531598

[27] Charles S. Peirce and Nathan Houser and Christian Kloesel, (eds.). 1878/1992. . Indiana University Press, Bloomington and Indianapolis, IN, Chapter The Doctrine of Chances, 142–154.

[28] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. Association for the Advancement of Artificial Intelligence, Palo Alto, CA, 8732–8740.

[29] Alana Semuels. 2018. The Internet is Enabling a New Kind of Poorly Paid Hell. *The Atlantic* (jan. 2018). https://www.theatlantic.com/business/archive/2018/01/amazon-mechanical-turk/551192/

[30] M. Six Silberman, Lilly Irani, and Joel Ross. 2010. Ethics and Tactics of Professional Crowdwork. *XRDS* 17, 2 (Dec. 2010), 39–43. https://doi.org/10.1145/1869086.1869100

[31] M. S. Silberman, B. Tomlinson, R. LaPlante, J. Ross, L. Irani, and A. Zaldivar. 2018. Viewpoint: Responsible research with crowds: pay crowdworkers at least minimum wage. *Commun. ACM* 61, 3 (2018), 39–41. https://doi.org/10.1145/3180492

[32] Robyn Speer, Joshua Chin, and Catherine Havasi. 2018. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. arXiv:1612.03975 [cs.CL]

[33] Richard Stallman. 1993. The GNU Manifesto. https://www.gnu.org/gnu/manifesto.html

[34] Luis Suarez-Villa. 2009. *Technocapitalism : A Critical Perspective on Technological Innovation and Corporatism.* Temple University Press, Philadelphia, PA.

[35] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. arXiv:1811.00937 [cs.CL]

[36] Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the First Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain, 53–59.

[37] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. *Discriminating Systems: Gender, Race, and Power in AI.* Technical Report. AI Now Institute. https://ainowinstitute.org/discriminatingsystems.pdf

[38] Mark E. Whiting, Grant Hugh, and Michael S. Bernstein. 2019. Fair Work: Crowd Work Minimum Wage with One Line of Code. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 197–206. https://ojs.aaai.org/index.php/HCOMP/article/view/5283