Issue 3

PARROTS

Are Large Language Models Our Limit Case?

Lauren Klein 🔗

How is it possible to go forward with the development and use of large language models, given the clear evidence of just how biased, how incomplete, and how harmful — to both people and the planet — these models truly are? This is the question that "Stochastic Parrots" productively sets in motion, and that has continued to reverberate throughout the artificial intelligence (AI) and machine learning (ML) community since the paper's release.[1]

This is also a question — or, rather, a more specific version of a question — that I've been thinking through for a number of years. In *Data Feminism*, for example, Catherine D'Ignazio and I ask more broadly how datasets and data systems, which so often encode and amplify power differentials, might instead be reimagined so that they can challenge and rebalance those differentials.[2] We document myriad instances of biased and incomplete datasets, as well as harmful and oppressive data systems that were conceived without attention to — let alone the involvement of — the communities most impacted by those systems.[3] We also consider the environmental and economic costs of the computing infrastructure required to support such systems.[4] Yet we maintain a guarded optimism that data science, if intentionally deployed, can be used to challenge unjust power systems. Throughout the book, we advance a view that it still may be possible to remake the field of data science by building coalitions across communities and taking differential power into account, so as to wield the power of data with care and in the service of justice. The seven principles of data feminism that we describe in the book — examine power, challenge power, rethink binaries and hierarchies, elevate emotion and embodiment, embrace pluralism, consider context, and make labor visible — were intended to structure this transformative work.[5]

But when "Stochastic Parrots" was released, and when Dr. Gebru and Dr. Mitchell were subsequently fired — and then subjected to so much harassment and defamation online — it called into question the degree to which the transformative data science that Catherine and I had described in the book was truly possible.[6] With the opportunity to reengage with "Stochastic Parrots," as this roundtable has invited us to do, I've returned to this question of degree, which — one year later — now seems more accurately described as a question of bounds: Are large language models (LLMs) our limit case? In other words, are LLMs simply too big, and too ethically, environmentally, and epistemologically compromised, for humanities scholars to abide?

> *Are LLMs simply too big, and too ethically, environmentally, and epistemologically compromised, for humanities scholars to abide?*

I will admit that I do not yet have a definitive answer to this question. But as a way of working through my current thinking, I will offer three unequivocal assertions:

1. There is no outside of unequal power.
2. All technologies are imbricated in this unequal power.
3. Refusal is, in itself, a generative act.
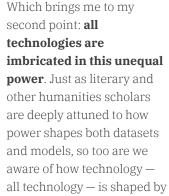
In what follows, I elaborate each of these points.

〵

First, there is no outside of unequal power. This is why an attention to power matters so much for discussions of data and models: it overdetermines not only the data we can collect and the datasets we can access, but also the research questions that we can explore. In *Data Feminism*, Catherine and I demonstrate how the financial and computational resources required to collect and analyze data at scale result in efforts most often undertaken by large corporations (and other well-resourced institutions) for their own profit and benefit, and at the expense of everyone else. This is why, to take one famous example that we discuss in the book, Target is able to analyze its own customer data in order to predict whether or not a person is pregnant — and then use that same data to sell them baby products; but there is not enough actual medical data that can be analyzed in order to predict whether that same pregnant person will be at risk of dying in childbirth.[7] Or, to take an example discussed in "Stochastic Parrots," why the Colossal Clean Crawled Corpus, even though it is less explicitly sexist, racist, and xenophobic than the Common Crawl corpus from which it is derived, remains sexist and racist and xenophobic in more subtle ways: because it was created by filtering out documents from the Common Crawl corpus that contained one or more words drawn from a list of "Dirty, Naughty, Obscene, or Otherwise Bad Words," it also excludes all discussions *about* those words — including direct critiques of those words or, in certain cases, discussions by those who might want to reclaim them.[8] More recent research by Gururangan et al. has shown that even more sophisticated quality filters, such as the classifier employed to cull the dataset on which GPT-3 is trained, demonstrate significant stylistic as well as thematic preferences, preferences which correlate with a number of proxies for high socioeconomic status.[9]

Literary and historical corpora are not exempt from the influence of unequal power. We all know (or should know) that the HathiTrust corpus, because it is drawn from the collections of major research libraries, reflects the collecting preferences of those libraries (and the preferences of several libraries in particular) rather than literary production writ large.[10] More domain-specific corpora, which would seem to avoid some of these issues as a result of their narrower scope, nonetheless continue to fall prey to the politics of digitization, as Katherine Bode and others have astutely observed.[11] And those of us who work on texts related to the history of slavery, in particular, have long had to reckon with the politics of the archive itself: the fact that the first-hand accounts that might offer us the most direct access to the lives of the enslaved scarcely exist at all is because of the historical structures of power that control not only what enters the archive but also who is authorized to even write. That there is no outside of this unequal power

is the starting point for our work. We must build our datasets and our models with our eyes wide open to this fact. As we acknowledge what we might learn from any new analyses, we must continue to account for the archive's "null values," as Jessica Marie Johnson describes them: spaces held open for the people and stories that we know to have existed in their own time, even as they remain unknown to us in the present.[12]

> *Academic researchers, especially at public institutions, are functionally if not ideologically compelled to concede to this asymmetrical configuration of power and resources.*

Which brings me to my second point: **all technologies are imbricated in this unequal power**. Just as literary and other humanities scholars are deeply attuned to how power shapes both datasets and models, so too are we aware of how technology — all technology — is shaped by power as well. One could point to broad histories of computing, statistics, or surveillance for evidence of this fact.[13] But we could also look to work that has explored the imbrications of power in specific natural language processing (NLP) and ML techniques themselves. Work by Jeff Binder, for example, has located the origins of topic modeling in the need for U.S. intelligence agencies to quickly scan international newswires for potential geopolitical conflicts.[14] Melanie Walsh, for another, has recently reminded us that the annotated OntoNotes corpus, on which spaCy's language models were trained, was developed with funding by the U.S. Department of Defense's Defense Advanced Research Projects Agency (DARPA) — as is so much of NLP/ML/AI work today.[15] This circles back to a point that Catherine and I make in *Data Feminism*, about unequal power being the result of unequal resources. This point is made in "Stochastic Parrots" as well: because these technologies are so resource-dependent, not only in terms of energy, cost, and computing power, but also because of very specific technical expertise, they are necessarily developed by people at elite and well-resourced institutions who are rarely required — either by inclination or circumstance — to take power into account.

Meredith Whittaker, another former Google employee forced out because of her labor-organizing efforts there, observes as much in a recent essay that documents the "capture" of supervised machine learning algorithms (including large language models) by Big Tech.[16] Whittaker explains how any gains in performance they might achieve are the result not of any major algorithmic or architectural innovation, but rather, of what existing algorithms can do "when matched with large-scale data and computational resources."[17] These are resources that only big tech firms control. So when academic researchers seek to engage with these developments, they find themselves beholden to the very same tech firms for the data and computing infrastructure, and very often funding, as a necessary precondition for contributing on an equal plane. The systematic defunding of higher education that has taken place over the past several decades is not the focus of Whittaker's piece, but this issue is the other side of the same neoliberal coin. Academic researchers, especially at public institutions, are functionally if not ideologically compelled to concede to this asymmetrical configuration of power and resources as a result of the federal, state, and institutional policies that have transferred the responsibility of supporting and sustaining research (in terms of computing infrastructure, student support, and even their own salaries) to the scholars themselves. And all of this is to say nothing of corporations like Facebook that are actively and intentionally wielding their data and algorithms to retain their own power, even as they know full well the harms that their products produce.[18]

Given this ethical, intellectual, and economic double-bind, it's not surprising that the most clear and compelling response may be to refuse — to refuse to develop new models, to refuse to improve existing ones, or even to refuse to participate in this work altogether. This brings me to my third assertion: that **refusal is, in itself, a generative act**. I've long admired the work of Dr. Joy Buolamwini and the Algorithmic Justice League (AJL), for example. This work began (in work coauthored with Dr. Gebru) by identifying biases in the image datasets used to train three major gender classification software libraries.[19] But when the initial audit found significant error rates in how the software classified images of women, and images of dark-skinned women in particular, Buolamwini's response was *not* to suggest that the training data be "debiased" or otherwise improved.[20] Rather, because she recognized how improved gender classification software (and facial recognition software more generally) would most likely be used to increase the policing and surveillance of Black and brown people, she used the evidence of her paper with Gebru to instead call for a ban on facial recognition software altogether.[21]

In terms of text analysis tools more specifically, I've followed with interest the recent actions by the team behind ml5.js, the JavaScript-based machine learning library, which upon discovering racist language in its sample word2vec model, around which its documentation is based, decided to remove it (and as a result, render the entire library nonfunctional) until an alternate model could be trained.[22] In both of these cases, the AJL and ml5.js, we see evidence of how refusal — especially when accompanied (as feminists advise) by a recommitment to values, or action, or both — can clear the space to imagine alternate possibilities.[23]

"Stochastic Parrots" participates in this reimagining by describing an alternative approach to technical research, one in which issues of cost, access, potential harms, and potential benefits are addressed early in the research process. This slower and more intentional process also allows for input from — and, ideally, meaningful collaboration with — impacted communities. This process echoes some of what Catherine and I have described as data science for good vs. data science for co-liberation: the latter imagines a way of doing data science in which those from both dominant and minoritized groups work together to free themselves from the oppressive systems that harm all of us.[24]

	"Data for good"	Data for co-liberation
Leadership by members of minoritized groups working in community		✓
Money and resources managed by members of minoritized groups		✓
Data owned and governed by the community		✓
Quantitative data analysis "ground truthed" through a participatory, community-centered data analysis process		✓
Data scientists are not rock stars and wizards, but rather facilitators and guides		✓
Data education and knowledge transfer are part of the project design		✓
Building social infrastructure—community solidarity and shared understanding—is part of the project design		✓

Above: Features of 'data for good' versus data for co-liberation, from Catherine D'Ignazio and Lauren F. Klein, Data Feminism (MIT Press, 2020), p. 140.

But the discussion of LLMs at the center of "Stochastic Parrots" complicates this vision of data for co-liberation in necessary ways, because LLMs may well function as a limit case. This is for a number of reasons. For one, the computational resources required to train up larger and larger models may make it infrastructurally impossible to allow anyone from outside of these institutions — let alone members of minoritized groups — to assume primary leadership of these models' training or future development. The same holds for financial leadership, since LLMs are just as resource-intensive with respect to cost as they are to compute. It is difficult to see how any corporation — which, by definition, is driven by its own bottom line — would allow an outside group to independently manage a budget that reflected one of these models' true economic cost, even as that same corporation might bestow seemingly generous grants to outside groups for specific purposes.

Along with questions about resources are questions about the models themselves. As the authors of "Stochastic Parrots" spell out, the predictive power of LLMs derives, in large part, from the size of the datasets used to train them. Is it possible for a single community, or even a consortium of impacted groups, to own and govern the increasing amount of training data that is required to train a new model from the ground up? Furthermore, without explainability mechanisms co-designed by communities, rather than by computer scientists alone, how might

the results of any particular LLM-based data analysis project be ground-truthed by the community members themselves?

There are additional ethical questions that arise as a result of this range of dependencies. Is it possible for corporations and their own technical workers to participate in a process of knowledge transfer while remaining uncompromised by their profit-driven agenda? How can academic researchers, if brought into this process, ensure that their own values remain uncompromised? What alternative infrastructures — social, technical, financial, or governmental — must be imagined such that Big Tech does not remain a required research partner? And even if removed from the corporate sphere, can LLMs — defined as they are by their size and scale — ever be enlisted in the necessarily slow, careful, and localized work of building solidarity and shared understanding?

Quantitative literary and cultural studies scholars can continue to learn from the work that the authors of "Stochastic Parrots" have undertaken since the paper's release (and even since the roundtable which prompted these remarks), which addresses many of these issues head on. For example, in December 2021, Dr. Gebru announced the launch of DAIR, the Distributed AI Research Institute, which is guided by a commitment to "ensur[ing] that researchers from many different backgrounds can participate while embedded in their communities."[25] Just a few months earlier, in August 2021, Dr. Mitchell announced that she would be joining Hugging Face, a startup that is working to provide open-source alternatives to the language models produced by the Big Five, as well as software libraries which increase access to existing models and examples of documentation and other best practices.[26] Meanwhile, McMillan-Major has continued her work on best practices for documenting the datasets employed in NLP research,[27] while Dr. Bender has continued to draw attention to the limits of LLMs as well as their potential harms, both through publications aimed at the AI, ML, and NLP research communities and in comments addressed to the public at large.[28]

How can we reconcile the historical specificity that we so value in our own research with the fact that even the most appropriate LLM for historical scholarship may be trained on data so temporally distant from the time period that bounds our own scholarly expertise?

But certain questions that pertain to humanistic inquiry in particular, both methodological and epistemological, remain to be addressed by quantitative literary and cultural studies scholars. For example, even as there begin to exist LLMs that are trained on historical corpora, the amount of data that is required results in training datasets with timespans — 1450 to 1950, in the case of MacBERTh — that far exceed any disciplinary sense of periodization.[29] How can we reconcile the historical specificity that we so value in our own research with the fact that even the most appropriate LLM for historical scholarship may be trained on data so temporally distant from the time period that bounds our own scholarly expertise? Furthermore, even as we know to fine-tune such a model on our own more curated datasets, how are we to measure the effects of that fine-tuning in ways that are meaningful to us as humanities scholars? When parameters no longer correspond to specific textual or linguistic features, as with earlier model architectures, we will require even more creative ways to understand the significance of the

texts contained in our curated datasets in relation to those on which the larger model was trained.

In addition, we must consider how decades of feminist thinking — and, for that matter, much of the most profound of humanities scholarship — has confirmed how a single voice at the margins can tell us just as much as (if not more than) a large group at the center. How do we hold fast to this fact as the allure of LLMs, enlisted as they are in the service of "shared" or generalizable tasks, continues to mount? How can we envision methods to engage these models in ways that center marginalized voices and the texts that document them? How can we amplify rather than merely assimilate the important oppositional ideas that these texts record? And how can we do so while remaining mindful of the ideas — and the people behind them — that these models cannot or at times should not subsume?

The answers to these questions we might, in turn, bring back to the authors of "Stochastic Parrots," augmenting their vision of the reimagining of large language models and their required infrastructures that must necessarily take place. As humanities scholars, we also must recommit to showing how literary, cultural, and historical context not only enriches our present understanding of LLMs, but is required for all future model-based research. After all, this is the set of contexts from which large language models emerged, and it is only with a deep knowledge of these contexts that we can fully understand their uses and limits.

1. Emily M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ," *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York: Association for Computing Machinery, 2021), 610–23, https://doi.org/10.1145/3442188.3445922. On "community" as an empty signifier, especially in the AI ethics space, see J. Khadijah Abdurahman, "Holding to Account: Safiya Umoja Noble and Meredith Whittaker on Duties of Care and Resistance to Big Tech," *Logic Magazine*, Dec. 2021, https://logicmag.io/beacons/holding-to-account-safiya-umoja-noble-and-meredith-whittaker/. ↵

2. Catherine D'Ignazio and Lauren F. Klein, *Data Feminism* (Cambridge: MIT Press, 2020), https://data-feminism.mitpress.mit.edu/. ↵

3. See D'Ignazio and Klein, *Data Feminism*, chap. 1, https://data-feminism.mitpress.mit.edu/pub/vi8obxh7/release/4; and chap. 5, https://data-feminism.mitpress.mit.edu/pub/2wu7aft8/release/3. ↵

4. See D'Ignazio and Klein, *Data Feminism*, chap. 1, https://data-feminism.mitpress.mit.edu/pub/vi8obxh7/release/4 ; and chap. 7, https://data-feminism.mitpress.mit.edu/pub/0vgzaln4/release/3/. ↵

5. See D'Ignazio and Klein, introduction to *Data Feminism*, https://data-feminism.mitpress.mit.edu/pub/frfa9szd/release/6. ↵

6. On Dr. Gebru's firing, see Cade Metz and Daisuke Wakabayashi, "Google Researcher Says She was Fired Over Paper Highlighting Bias In A.I.," *New York Times*, Dec. 3, 2020, https://www.nytimes.com/2020/12/03/technology/google-researcher-timnit-gebru.html; on Dr. Mitchell's firing, see Cade Metz, "A Second Google A.I. Researcher Says the Company Fired Her," *New York Times*, Feb. 19, 2021, https://www.nytimes.com/2021/02/19/technology/google-ethical-artificial-intelligence-team.html; on the harassment that followed, see Zoe Schiffer, "Timnit Gebru Was Fired from Google — Then the Harassment Arrived," *The Verge*, Mar. 5, 2021, https://www.theverge.com/22309962/timnit-gebru-google-harassment-campaign-jeff-dean. For Catherine's response to Dr. Gebru's firing in particular, see Katlyn Turner, Danielle Wood, and Catherine D'Ignazio, "The Abuse and Misogynoir Playbook," in *The State of AI Ethics Report* (Montreal AI Ethics Institute, 2021), 14–34, https://montrealethics.ai/wp-content/uploads/2021/01/The-State-of-AI-Ethics-Report-January-2021.pdf#page=15 . ↵

7. D'Ignazio and Klein, *Data Feminism*, chap. 1, https://data-feminism.mitpress.mit.edu/pub/vi8obxh7/release/4. ↵

8. Bender et al., "Stochastic Parrots," 614. ↵

9. Suchin Gururangan et al., "Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection", arXiv, updated Jan. 26, 2022, https://arxiv.org/abs/2201.10474. ↵

10. Benjamin Schmidt, "What's in the HathiTrust?" *Sapping Attention* (blog), Feb. 3, 2022, https://sappingattention.blogspot.com/2019/03/whats-in-hathi-trust.html. ↵

11. Katherine Bode, "Why You Can't Model Away Bias," *Modern Language Quarterly* 81, no. 1 (2020): 95–124. For an example and analysis of the politics of digitization in action, see Benjamin Fagan, "Chronicling White America," *American Periodicals* 26, no. 1 (2016): 10–13. ↵

12.   Jessica Marie Johnson, *Wicked Flesh: Black Women, Intimacy, and Freedom in the Atlantic World* (Philadelphia: Univ. of Pennsylvania Press, 2020). ↵

13.   See Simone Browne, *Dark Matters: On the Surveillance of Blackness* (Durham: Duke Univ. Press, 2015); Mar Hicks, *Programmed Inequality: How Britain Discarded Women Technologists and Lost Its Edge in Computing* (Cambridge: MIT Press, 2017); and Banu Subramaniam, *Ghost Stories for Darwin: The Science of Variation and the Politics of Diversity* (Urbana: Univ. of Illinois Press, 2014). ↵

14.   Jeffrey M. Binder, "Alien Reading: Text Mining, Language Standardization, and the Humanities," in *Debates in the Digital Humanities 2016*, ed. Matthew K. Gold and Lauren F. Klein (Minneapolis: Univ. of Minnesota Press, 2016), https://dhdebates.gc.cuny.edu/read/untitled/section/4b276a04-c110-4cba-b93d-4ded8fcfafc9. ↵

15.   Melanie Walsh, *Introduction to Cultural Analytics and Python,* version 1.1.0 (Aug. 31, 2021), https://doi.org/10.5281/zenodo.4411250. ↵

16.   On retaliation against Whittaker, see Nitasha Tiku, "Google Walkout Organizers Say They're Facing Retaliation," *Wired*, Apr. 2, 2019, https://www.wired.com/story/google-walkout-organizers-say-theyre-facing-retaliation/. ↵

17.   Meredith Whittaker, "The Steep Cost of Capture," *ACM Interactions,* vol. 28, no .6 (Nov.–Dec. 2021): 50, https://interactions.acm.org/archive/view/november-december-2021/the-steep-cost-of-capture. ↵

18.   See, for example, Karen Hao, "The Facebook Whistleblower Says Its Algorithms Are Dangerous. Here's Why," *MIT Technology Review,* Oct. 5, 2021, https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/; and Karen Hao, "She Risked Everything to Expose Facebook. Now She's Telling Her Story," *MIT Technology Review*, July 29, 2021, https://www.technologyreview.com/2021/07/29/1030260/facebook-whistleblower-sophie-zhang-global-political-manipulation/. ↵

19.   Joy Buolamwini and Timnit Gebru, "Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018): 77–91. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf ↵

20.   For the audit, see Joy Buolamwini and Timnit Gebru, "Gender Shades," accessed May 15, 2022, http://gendershades.org/index.html. ↵

21.   See "Policy/Advocacy," Algorithmic Justice League, accessed May 15, 2022, https://www.ajl.org/library/policy-advocacy. ↵

22.   See @ml5js, "The team recently learned that the example models used to demonstrate our word2vec functionality include racial slurs and other offensive words," Twitter, Oct. 6, 2021, 10:46 a.m., https://twitter.com/ml5js/status/1445762321444315147. ↵

23.   On feminist refusal, see Marika Cifor et al., "Feminist Data Manifest-No" (2019), https://www.manifestno.com/; and Patricia Garcia et al., "No: Critical Refusal as Feminist Data Practice," abstract, in *CSCW '20 Companion: Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing* (New York: Association for Computing Machinery, 2020), 199–202, https://doi.org/10.1145/3406865.3419014. ↵

24.   D'Ignazio and Klein, *Data Feminism*, chap. 5, https://data-feminism.mitpress.mit.edu/pub/2wu7aft8/release/3\#nobxi408tlj. ↵

25.   "Research Philosophy," Distributed Artificial Intelligence Research Institute, accessed May 15, 2022, https://www.dair-institute.org/research. ↵

26.   For reportage on Hugging Face's most recent round of funding, see, for example, Romain Dillet, "Hugging Face Reaches $2 Billion Valuation to Build the GitHub of Machine Learning," *TechCrunch*, May 9, 2022, https://techcrunch.com/2022/05/09/hugging-face-reaches-2-billion-valuation-to-build-the-github-of-machine-learning/. ↵

27.   See, for example, Angela McMillan-Major, Emily Bender, and Batya Friedman, "Data Statements: Documenting the Datasets Used for Training and Testing Natural Language Processing Systems" (poster, Scholarly Communication in Linguistics: Resource Workshop and Poster Session, Linguistic Society of America, virtual, Jan. 6, 2022), https://www.linguisticsociety.org/system/files/abstracts/summary/Scholarly%20Communication%20in%20Linguistics.pdf. ↵

28.   See, for example, Emily M. Bender, "On NYT Magazine on AI: Resist the Urge to Be Impressed," *Medium* (blog), Apr. 17, 2022, https://medium.com/@emilymenonbender/on-nyt-magazine-on-ai-resist-the-urge-to-be-impressed-3d92fd9a0edd; Steven Johnson, "A.I. is Mastering Language. Should We Trust What It Says?" *New York Times Magazine*, Apr. 17, 2022, https://www.nytimes.com/2022/04/15/magazine/ai-language.html; and Inioluwa Deborah Raji et al., "AI and the Everything in the Whole Wide World Benchmark," *NeurIPS* 2021, https://arxiv.org/abs/2111.15366. ↵

29.   Enrique Manjavacas and Lauren Fonteyn, "MacBERTh: Development and Evaluation of a Historically Pre-trained Language Model for English (1450–1950), in *Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH)* (Stroudsburg: Association of Computing and the Humanities, 2021), 23–36, https://rootroo.com/downloads/nlp4dh_proceedings_draft.pdf. The project page currently lives at https://macberth.netlify.app/. ↵